



UNIVERSIDAD
DE MURCIA

Escuela
de Doctorado

TESIS DOCTORAL

*Hacia la interoperabilidad y nuevos enfoques
metodológicos para la evaluación escalable
basada en juegos*

*Towards Interoperability
and Novel Methodological
Approaches for Scalable
Game-Based Assessment*

AUTOR/A

Manuel Jesús Gómez Moratilla

DIRECTOR/ES

Félix Jesús García Clemente
José Antonio Ruipérez Valiente

2025



UNIVERSIDAD
DE MURCIA

Escuela
de Doctorado

TESIS DOCTORAL

*Hacia la interoperabilidad y nuevos enfoques
metodológicos para la evaluación escalable
basada en juegos*

*Towards Interoperability
and Novel Methodological
Approaches for Scalable
Game-Based Assessment*

AUTOR/A

Manuel Jesús Gómez Moratilla

DIRECTOR/ES

Félix Jesús García Clemente
José Antonio Ruipérez Valiente

2025



DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TÍTULO DE DOCTOR/A

Aprobado por la Comisión General de Doctorado el 19 de octubre de 2022.

Yo, D. Manuel Jesús Gómez Moratilla, habiendo cursado el Programa de Doctorado en Informática de la Escuela Internacional de Doctorado de la Universidad de Murcia (EIDUM), como autor/a de la tesis presentada para la obtención del título de Doctor/a titulada:

Towards Interoperability and Novel Methodological Approaches for Scalable Game-Based Assessment / Hacia la interoperabilidad y nuevos enfoques metodológicos para la evaluación escalable basada en juegos

y dirigida por:

D.: Félix Jesús García Clemente
D.: José Antonio Ruipérez Valiente

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como compendio de publicaciones, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

Murcia, a 19 de junio de 2025

D. Manuel Jesús Gómez Moratilla



Acknowledgements

With this writing, I bring to a close four years of experiences, learning, and growth, both professionally and personally. Thank you to all those who, in one way or another, have made this journey a little easier and so much more meaningful. This document should not bear only my name, but also the names of all of you who, without even realizing it, have helped write each and every line of this Ph.D. thesis.

Thank you all for being part of this journey.

Agradecimientos

Con esta escritura pongo fin a cuatro años de experiencias, aprendizajes y crecimiento, tanto en el ámbito profesional como en el personal. Gracias a todas aquellas personas que, de una forma u otra, han hecho este camino un poco más fácil y mucho más significativo. Este documento no debería llevar solo mi nombre, sino también el de todos vosotros que, sin saberlo, habéis contribuido a escribir cada una de las líneas de esta tesis doctoral.

Gracias a todos por formar parte de este camino.

Abstract

As the 21st century progresses, new assessment methodologies are emerging and challenging traditional approaches for evaluating knowledge, skills, and behaviors. Among these innovations, Game-Based Assessment (GBA) has been gaining increasing attention in recent years due to its potential to enhance current assessment practices. Although games have always been an integral part of human societies, they are now being increasingly explored as powerful tools for learning, skill development, and assessment. Specifically, Serious Games (SGs), which do not have entertainment, enjoyment, or fun as their primary purpose, have gained significant attention over the past decade. GBAs have been proved to be robust alternatives to conventional forms of assessment, such as paper-and-pencil tests, which often present individual and decontextualized items to learners, making it difficult to capture the complexity of certain skills and behaviors, as well as their application in real-world contexts. In contrast, GBAs offer realistic and authentic scenarios that support the contextualized application of knowledge and skills.

GBAs are suitable for use in different contexts and environments, including education, medical settings, and professional environments for purposes such as employee selection and training. As a result, large amounts of data are generated, offering unprecedented opportunities for scientific exploration and technological progress. These data can be leveraged in many ways, and recent advancements in Artificial Intelligence (AI) models and algorithms have paved the way for creating even more sophisticated assessment systems. However, there is still a set of limitations to be addressed before the full potential of GBAs can be realized in real-world applications. Previous research often reports methodological limitations, as the literature typically relies on basic metrics and indicators for assessment that lack the depth and complexity required for capturing certain skills and behavioral patterns. Moreover, educators frequently report a lack of guidance on how to integrate these type of assessments into their teaching practices, challenging their practical adoption. The addition of assessment features in games is often seen as a costly process, since the assessment machinery is typically designed specifically for each individual game. Finally, technical limitations are also commonly reported, as the large amounts of data being generated by GBAs require scalable architectures and efficient data processing pipelines.

Given the significant potential of GBAs to transform assessment practices, the main purpose of this Ph.D. thesis is to address both the methodological and technical aspects of the field, with a focus on exploring and advancing how GBAs can serve as an innovative solution for delivering valid, meaningful, and adaptive assessment

experiences. Specifically, this dissertation poses five objectives:

The first proposed objective concerns a comprehensive examination and evaluation of GBA field, particularly with the recent rise in the popularity of games and digital learning. To the best of our knowledge, there is no existing research providing an in-depth analysis of the current state of the field. Thus, this objective involves performing a systematic analysis of the GBA developments, techniques and tools, highlighting its strengths and weaknesses. Moreover, this thesis proposes potential directions in which GBA can be further advanced and applied by addressing current limitations.

The second objective of this Ph.D. thesis focuses on designing and developing an interoperable semantic model for log data. Although the use of games in educational and professional settings offers new opportunities to analyze and evaluate learners' behavior and performance, the lack of standardization in this area limits interoperability and reproducibility, making it difficult to generalize findings and reuse assessment solutions. This objective, therefore, entails the creation of an interoperable model capable of integrating log data from a wide variety of games into a unified knowledge structure. Additionally, it aims to validate the proposed model using widely accepted metrics drawn from existing literature.

The third objective of this dissertation is geared towards the creation of an efficient architecture capable of performing GBAs at scale. Given the vast amount of data generated by learners and the computational demands of modern techniques and algorithms, there is a clear need for a scalable and distributed architecture that can analyze large volumes of log data in real time. With this objective in mind, the thesis seeks to design a Big Data architecture that supports interoperability, efficient analysis, and robust integration with AI models.

The fourth proposed objective aims to enhance student performance prediction by using Explainable AI (XAI). Despite recent advances in AI techniques, explainability remains a significant challenge, particularly in contexts where stakeholders are typically non-technical users who require clear and interpretable insights to understand the output of such techniques. Through this objective, the thesis aims to establish a comprehensive framework for interpretable models in GBAs, facilitating a clearer understanding of AI-driven assessments.

Last but not least, the fifth objective of this thesis seeks to optimize the data labeling process for AI techniques in the GBA domain through the creation of a practical tool. The accuracy and quality of human-labeled data is an essential part of building reliable AI models, as it directly influences the performance and validity of the results. However, the data labeling process is often seen as costly and time-consuming, and researchers frequently rely on rudimentary methods such as Excel worksheets. To address this gap, the final objective focuses on the design and development of a web-based tool aimed at improving both the efficiency and accuracy of the data labeling process within GBA environments.

To achieve the proposed objectives, a clearly structured methodology was followed, employing a scientific approach grounded in the continuous study of the state of the art and the analysis of results obtained throughout the different stages of this research. First, a detailed analysis of recent studies in the GBA field addresses the first objective. This analysis followed a standard systematic review methodology

based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The systematic review facilitated the identification of current trends, commonly used methodologies, and potential gaps in the existing literature. Secondly, to address the interoperability issue identified in the literature review, we designed and implemented an ontology-driven model aimed at standardizing various log data formats from different games into a unified model. Following the development of this ontology, a formal evaluation and real-world testing with actual data were conducted, therefore fulfilling the second objective. Then, to address the third objective and provide a scalable architecture in which the ontology model could be integrated, an interoperable framework was developed using an open-source structured data processing engine as the basis to enable computation over large-scale datasets. Subsequently, a case study validation was carried out to demonstrate how the architecture could be applied in real-world scenarios.

The next step of the methodology fulfills the fourth objective and involved the development of a learner performance prediction model for a SG using Machine Learning (ML) techniques. Then, interpretability was achieved by considering whether the selected model was inherently explainable. If it was not, XAI techniques were applied to generate meaningful explanations for the model's predictions. Finally, the fifth step of the methodology consisted of developing a practical tool specifically designed for labeling various types of GBA data, with support for *audio*, *video*, and *game-event* inputs. This comprehensive methodology resulted in the publication of five peer-reviewed scientific articles, which collectively form the basis of this Ph.D. thesis. It should be noted that the Ph.D. candidate served as the first author of each of these publications, all of which are included in the body of this document.

As a result of the first step of the methodology, the systematic review of literature analyzed 65 research papers published between 2013 and 2020. The analysis highlighted the relevance of GBAs in *K-16 education* (most commonly in high and middle school), the *workplace*, and *medical settings*. Furthermore, it revealed that *descriptive statistics* was the main analytical method used, while *ML* and *Deep Learning (DL)* methods were applied in only a minority of studies. Additional findings included the main research purposes, data availability, and domain categories represented across the reviewed papers, among other aspects. Finally, the review identified key open challenges in the field, such as issues related to replication due to the lack of transparency, transferability of research, and both methodological and technical limitations.

Secondly, the thesis introduced a novel ontology that conceptualizes the core concepts of the GBA field, such as *Game*, *Scenario*, or *Attempt*. Relationships between concepts enhance the model's expressiveness and enable executing more complex ontology queries to extract valuable knowledge from the processed data. To practically validate the ontology, the study employed both established metrics from previous literature and newly designed metrics, including *persistence* and *play styles*. Finally, a case study was presented to evaluate the interoperability and usability of the developed model using data from ten different SGs. Following the model creation, the third contribution of this dissertation is the development of a scalable and interoperable architecture. The proposed framework has five main components: a *prepro-*

cessing module that transforms raw game data into an ontology-compatible RDF format; an *analytics, inference, and querying module* that processes RDF triples as distributed data structures; a *metrics module* that computes assessment metrics using SPARQL and allows results to be exported in various formats such as plain text or CSV; an *authentication and authorization module* that manages access control based on predefined user roles; and a *Service API*, which enables external access to the framework and introduces the paradigm of Game-Based Assessment as a Service (GBaaS). A performance evaluation was conducted using different cluster configurations and dataset sizes. The best performing setup, consisting of one master node and four worker nodes, successfully processed two million user events (equivalent to data generated by 39 classrooms using a game for one hour per week over the course of a month) in an average of 107.2 minutes. These results demonstrate the efficiency of the proposed system in handling large-scale data, and confirms its potential for real-world applications.

Regarding the fourth objective, the next step in the dissertation involved the development of a real-time performance prediction model within *Shadowspect*, a geometry SG. Predictions were made at three different intervals (25, 50 and 75% of the average level completion time), with the Random Forest (RF) model achieving a balanced accuracy of 0.76 at the 25% interval, 0.772 at 50%, and 0.795 at 75%. This indicates strong predicting performance even in the early stages of gameplay. Since RF is considered a “black box” algorithm, the SHAP method was employed to explain individual predictions. This XAI technique enabled the identification of the most relevant factors contributing to students’ success in solving different levels. Finally, this result included a use case to demonstrate how interpretable models can be applied in the classroom to support individual students during gameplay.

The final contribution of this thesis presented a web-based tool, built using the *Django* framework, specifically designed to optimize the data labeling process in GBA environments. With this objective in mind, the tool incorporates a custom parser that transforms raw data into structured instances stored in the database, along with a *feature computing* module that analyzes the data and automatically calculates a set of features to provide context and support the labeling process. In addition, the tool offers several data visualization options for different data types: templates compatible with Unity WebGL for in-game replays, text-based replays, a video viewer for video recordings, and audio waveform visualizations for audio files. All these components form a comprehensive tool designed to address a common challenge in the field and optimize data labeling tasks, which are crucial for training reliable AI models in this area.

As main conclusions, this Ph.D. thesis underlines the value and potential of GBA for education and training, particularly emphasizing its potential to support the development and sustainability of 21-st century skills. These skills, such as collaboration or critical thinking, are increasingly considered essential in a rapidly transforming, technology-driven society. Although we have seen that valuable knowledge can be extracted and inferred from user interaction data, there remains a need for more standardized assessment frameworks, as the current specificity limits the replication of experiments and the transferability of results into practice. Moreover, the conducted review also revealed the importance of making good game designs in-

stead of relying on vague implementations, such as embedding hidden questionnaires within the gameplay.

It is also worth mentioning the need for more sophisticated assessment methods. Although the comprehensive datasets produced hold tremendous potential for applying novel techniques, researchers typically use simple metrics and indicators that fail to capture the complexity and context-dependence nature of certain skills and behaviors, such as the aforementioned 21-st century skills. However, for these advanced methods to perform well, large data sets are required, and with many studies in the literature reporting data sample size limitations, this remains a significant challenge in the field. Moreover, many of these emerging techniques are seen as humanly inexplicable by non-technical stakeholders, which also limits their applicability in real-world environments. Therefore, it is essential to continue exploring explainability techniques to provide these stakeholders with a clear understanding of how such techniques work, fostering trust, transparency, and informed decision-making in practical settings.

As part of the future work, this thesis proposes four different avenues to explore as a continuation of this research: first, the development of frameworks for the scalable design and integration of GBAs, with an emphasis on aligning game mechanics with learning objectives, assessment methods, and existing curricula to facilitate adoption in educational environments; second, the empirical validation of GBA solutions in real-world settings, focusing on the evaluation of assessment validity, usability, and accessibility to strengthen the proposed interoperable framework; third, the advancement of Human-in-the-Loop (HITL) approaches by leveraging the practical labeling tool to develop more sophisticated assessment models capable of capturing complex learner behaviors and supporting the training of reliable AI systems; and fourth, the integration of multimodal data into current assessment approaches to enrich models by converging evidence from multiple sources of data.

Resumen

A medida que el siglo XXI progresa, surgen nuevas metodologías de evaluación que desafían los enfoques tradicionales para evaluar conocimientos, habilidades y comportamientos. Entre estas innovaciones, la Evaluación Basada en Juegos (Game-Based Assessment, GBA) ha ido ganando atención durante los últimos años debido a su potencial para mejorar los enfoques de evaluación actuales. Aunque los juegos siempre han formado parte de las sociedades humanas, actualmente se exploran cada vez más como herramientas poderosas para el aprendizaje, el desarrollo de habilidades y la evaluación. Específicamente, los Juegos Serios (Serious Games), que no tienen como propósito principal el entretenimiento, la diversión o el ocio, han ganado la atención de muchas personas durante la última década. Las evaluaciones mediante juegos han demostrado ser muy buenas alternativas a las formas convencionales de evaluación como las pruebas escritas, que suelen presentar elementos individuales y descontextualizados. Este tipo de pruebas dificulta la captura de ciertas habilidades y comportamientos debido a su complejidad, y también su aplicación en contextos del mundo real. En cambio, las GBA ofrecen escenarios realistas y auténticos que favorecen la aplicación contextualizada del conocimiento y las habilidades de los usuarios.

Las GBA son adecuadas para su uso en distintos contextos y entornos, incluidos el ámbito educativo, los entornos médicos y los entornos profesionales, con fines como la selección o la formación y entrenamiento de personal. Como resultado, se generan una gran cantidad de datos, los cuales ofrecen oportunidades sin precedentes para la exploración científica y el progreso tecnológico. Estos datos pueden ser aprovechados de distintas formas, y los recientes avances en modelos y algoritmos de Inteligencia Artificial (IA) han abierto el camino para la creación de sistemas de evaluación todavía más sofisticados. Sin embargo, aún existen una serie de limitaciones que deben abordarse antes de que las GBA puedan desplegar todo su potencial en aplicaciones del mundo real. El estado del arte reporta a menudo limitaciones metodológicas, ya que la literatura suele basarse en métricas e indicadores básicos que carecen de la profundidad y complejidad necesarias para capturar ciertas habilidades y patrones de comportamiento. Por otra parte, los docentes suelen señalar la falta de orientación sobre cómo integrar este tipo de evaluaciones en sus prácticas pedagógicas, lo que dificulta su adopción práctica. Además, la incorporación de mecanismos de evaluación en los juegos suele considerarse un proceso costoso, ya que generalmente se diseñan sistemas de evaluación de forma totalmente específica para cada juego. Finalmente, también se suelen reportar de forma frecuente limitaciones técnicas, ya que el gran volumen de datos generados por las GBA requiere de arquitecturas

escalables y enfoques de procesamiento de datos eficientes.

Dado el gran potencial que tienen las GBA para transformar las prácticas de evaluación tradicionales, el principal objetivo de esta tesis doctoral es abordar tanto los aspectos metodológicos como los técnicos en el área, con un enfoque centrado en explorar y avanzar en cómo las GBA pueden servir como una solución novedosa e innovadora para proporcionar experiencias de evaluación válidas, robustas, y significativas. Específicamente, esta tesis plantea cinco objetivos:

El primer objetivo propuesto se refiere a un análisis y evaluación exhaustivos del campo de GBA, especialmente teniendo en cuenta el reciente aumento en la popularidad de los juegos y el aprendizaje digital. Hasta donde sabemos, no existe una investigación previa que ofrezca un análisis profundo del estado actual del campo. Por tanto, este objetivo implica realizar una revisión sistemática de los desarrollos, técnicas y herramientas en GBA, destacando sus fortalezas y debilidades. Además, la tesis también propone posibles direcciones de trabajo futuro para avanzar y aplicar evaluaciones basadas en juegos abordando las limitaciones actuales.

El segundo objetivo de esta tesis se centra en el diseño y desarrollo de un modelo semántico interoperable para datos de interacción de estudiantes con juegos. Aunque el uso de juegos en entornos educativos y profesionales ofrece nuevas oportunidades para analizar y evaluar el comportamiento y rendimiento de los estudiantes, la falta de estandarización en este ámbito limita la interoperabilidad y la reproducibilidad, lo que dificulta la generalización y reusabilidad de los resultados de investigación. Por lo tanto, este objetivo implica la creación de un modelo interoperable capaz de integrar datos de una amplia variedad de juegos en una estructura de datos unificada. Además, se propone validar el modelo utilizando métricas ampliamente aceptadas en la literatura existente.

El tercer objetivo de esta tesis se orienta al diseño y creación de una arquitectura eficiente capaz de ejecutar GBA a gran escala. Dada la gran cantidad de datos generados por los estudiantes y las exigencias computacionales de las técnicas y algoritmos actuales, existe una necesidad evidente de crear una arquitectura distribuida y escalable que permita analizar grandes cantidades de datos en un tiempo razonable. Con este propósito, la tesis busca diseñar una arquitectura “Big Data” que soporte la interoperabilidad, el análisis eficiente y la integración robusta con modelos de IA.

El cuarto objetivo propuesto apunta a mejorar la predicción del rendimiento estudiantil mediante el uso de Inteligencia Artificial Explicable (XAI). A pesar de los recientes avances en técnicas de IA, la explicabilidad sigue siendo un desafío importante, especialmente en contextos donde los actores involucrados suelen ser usuarios no técnicos que requieren información clara e interpretable para comprender los resultados de estas técnicas. A través de este objetivo, la tesis busca establecer un marco integral para modelos interpretables en GBA, facilitando una mejor comprensión de las evaluaciones impulsadas por IA.

Por último, el quinto objetivo de esta tesis se centra en la optimización del proceso de etiquetado de datos para técnicas de IA en el ámbito de las GBA mediante la creación de una herramienta práctica de etiquetado. En este sentido, la precisión y calidad de los datos etiquetados por humanos son fundamentales para construir modelos de IA robustos y confiables, ya que influyen directamente en el rendimiento

y la validez de los resultados. Sin embargo, los investigadores suelen ver el proceso de etiquetado como algo lento y costoso, y suelen tender a utilizar métodos rudimentarios como hojas de cálculo en Excel o anotaciones manuales. Para abordar este desafío, el último objetivo de esta tesis se enfoca en el diseño y desarrollo de una herramienta web destinada a mejorar tanto la eficiencia como la precisión del proceso de etiquetado en entornos de GBA.

Para alcanzar los objetivos propuestos, se ha seguido una metodología estructurada basada en un enfoque científico fundamentado en el estudio continuo del estado del arte y en el análisis de los resultados obtenidos a lo largo de las distintas etapas de esta investigación. En primer lugar, una revisión detallada de estudios recientes en el campo de las GBA aborda el primer objetivo. Este análisis se llevó a cabo siguiendo una metodología de revisión sistemática basada en las directrices PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). La revisión permitió identificar tendencias actuales, metodologías comúnmente utilizadas y brechas potenciales en la literatura existente. En segundo lugar, para abordar el problema de interoperabilidad identificado en la revisión, se diseñó e implementó un modelo ontológico orientado a estandarizar distintos formatos de datos provenientes de diferentes juegos en un modelo unificado. Tras el desarrollo de esta ontología, se realizó una evaluación formal y pruebas en entornos reales con datos reales, cumpliendo así con el segundo objetivo. Después, para abordar el tercer objetivo y proporcionar una arquitectura escalable en la que pudiera integrarse el modelo ontológico, se desarrolló un sistema interoperable utilizando un motor de procesamiento de datos estructurados de código abierto como base, lo que permitió el procesamiento de grandes volúmenes de datos de forma eficiente. Posteriormente, se llevó a cabo una validación mediante un caso de estudio para demostrar cómo puede aplicarse esta arquitectura en escenarios del mundo real.

El siguiente paso de la metodología responde al cuarto objetivo e implicó el desarrollo de un modelo de predicción del rendimiento del estudiante en un juego serio utilizando técnicas de Aprendizaje Automático (Machine Learning, ML). Posteriormente, se incorporó la interpretabilidad del modelo considerando si este era inherentemente explicable; en caso contrario, se aplicaron técnicas de XAI para generar explicaciones claras y sencillas de las predicciones del modelo. Finalmente, el quinto paso de la metodología consistió en desarrollar una herramienta práctica diseñada específicamente para el etiquetado de distintos tipos de datos en entornos de GBA, con soporte para entradas de *audio*, *video* y *datos de interacción*. Esta metodología dio como resultado la publicación de cinco artículos científicos revisados por pares, que en conjunto constituyen la base de esta tesis doctoral. Cabe destacar que el doctorando actuó como primer autor en cada una de estas publicaciones, todas ellas incluidas en el cuerpo del presente documento.

Como resultado del primer paso de la metodología, la revisión sistemática de la literatura analizó 65 artículos de investigación publicados entre los años 2013 y 2020. El análisis destacó la relevancia de las GBA en la educación K-16 (principalmente en educación primaria y secundaria), el ámbito laboral y los entornos médicos. Asimismo, reveló que los principales métodos utilizados fueron las estadísticas descriptivas, mientras que las técnicas de ML y Aprendizaje Profundo (Deep Learning, DL) solo se aplicaron en una minoría de los estudios. Entre otros hallaz-

gos, se identificaron los principales propósitos de investigación, la disponibilidad de datos y las categorías temáticas abordadas en los artículos analizados. Finalmente, la revisión identificó retos clave que aún permanecen sin resolver en el campo, como la limitada replicación de resultados debido a la falta de transparencia, la escasa transferibilidad de los resultados y diversas limitaciones, tanto metodológicas como técnicas.

En segundo lugar, la tesis presentó una nueva ontología que conceptualiza los elementos fundamentales del campo de las GBA, como *Juego*, *Escenario* e *Intento*. Por otra parte, la ontología incluye una serie de relaciones entre conceptos que logran enriquecer la expresividad del modelo y permiten ejecutar consultas más complejas, lo que facilita la extracción de conocimiento aún más valioso de los datos ya procesados. Para validar esta ontología de manera práctica, el estudio empleó tanto métricas ya consolidadas en la literatura previa como otras nuevas diseñadas, como *persistencia* o *estilos de juego*. Finalmente, se presentó un caso de estudio para evaluar la interoperabilidad y usabilidad del modelo utilizando datos de diez juegos serios distintos.

Tras la creación del modelo, la tercera contribución de esta tesis es el desarrollo de una arquitectura escalable e interoperable. El sistema desarrollado incluye cinco componentes principales: un *módulo de preprocesamiento* que transforma los datos en crudo del juego en formato RDF compatible con la ontología; un *módulo de análisis, inferencia y consulta*, que procesa un formato de triplas RDF como estructuras de datos distribuidas; un *módulo de métricas*, que calcula los indicadores de evaluación mediante consultas SPARQL y permite exportar los resultados en distintos formatos como texto plano o CSV; un *módulo de autenticación y autorización*, que gestiona el control de acceso según los roles de usuario definidos; y finalmente una *API*, que permite el acceso externo al sistema e introduce el paradigma de Evaluación Basada en Juegos como Servicio (Game-Based Assessment as a Service, GBaaS). Además, se llevaron a cabo pruebas de rendimiento utilizando distintas configuraciones de clúster y tamaños de conjuntos de datos. La configuración más eficiente logró procesar dos millones de eventos de usuario (equivalentes a los datos generados por 39 aulas utilizando un juego durante una hora a la semana durante un mes completo) en un promedio de 107.2 minutos. Estos resultados demuestran la eficiencia del sistema propuesto para el procesamiento de datos a gran escala y confirman su aplicabilidad en escenarios reales.

Respecto al cuarto objetivo, la siguiente etapa de la tesis consistió en el desarrollo de un modelo de predicción en tiempo real del rendimiento estudiantil en *Shadowspect*, un juego serio orientado a la geometría. Se realizaron predicciones sobre si el estudiante sería capaz de resolver o no el nivel en cuestión en tres intervalos distintos (25%, 50% y 75% del tiempo medio de finalización de nivel), obteniendo el modelo Random Forest (RF) una precisión de 0.76 en el primer intervalo, 0.772 en el segundo y 0.795 en el tercero. Esto indica un buen rendimiento predictivo incluso en las primeras etapas del juego, cuando aún se dispone de información limitada. Dado que RF es considerado un algoritmo de tipo “caja negra”, se utilizó el método SHAP para explicar las predicciones individuales. Esta técnica de XAI permitió identificar los factores más relevantes que contribuyen al éxito de los estudiantes al resolver los distintos niveles. Finalmente, este resultado incluye también un caso de

uso que demuestra cómo los modelos interpretables pueden aplicarse en el aula para apoyar individualmente a los estudiantes durante las sesiones de juego y evaluación.

La última contribución de esta tesis presentó una herramienta web, desarrollada sobre el framework de *Django*, diseñada específicamente para optimizar el proceso de etiquetado de datos en entornos GBA. Con este objetivo en mente, la herramienta incorpora un módulo de procesamiento personalizado que transforma los datos en crudo en instancias estructuradas almacenadas en la base de datos, junto con un módulo de cálculo de características que analiza los datos y calcula automáticamente un conjunto de variables para proporcionar contexto y apoyar el proceso de etiquetado. Además, la herramienta ofrece distintas opciones de visualización de datos según el tipo de estos: plantillas compatibles con Unity WebGL para repeticiones haciendo uso del propio motor del juego, repeticiones basadas en una representación textual de las acciones del usuario, un visualizador de video para grabaciones, y visualizaciones de ondas para archivos de audio. Todos estos componentes forman una herramienta muy completa orientada a abordar una limitación común en el área y a optimizar las tareas de etiquetado de datos, fundamentales para el entrenamiento de modelos de IA precisos y confiables en este dominio.

Como principales conclusiones, esta tesis doctoral subraya el valor y potencial de las GBA para la educación y la formación, destacando especialmente su capacidad para fomentar el desarrollo y mantenimiento de las habilidades del siglo XXI. Estas habilidades, como la colaboración o el pensamiento crítico, son cada vez más consideradas esenciales en una sociedad en transformación constante y mediada por la tecnología. Aunque hemos visto que es posible extraer e inferir conocimiento valioso a partir de los datos de interacción de los usuarios, sigue siendo necesario avanzar hacia marcos de evaluación más estandarizados, ya que la actual especificidad limita la replicación de experimentos y la transferibilidad de los resultados a la práctica. Además, la revisión realizada también puso de manifiesto la importancia de diseñar juegos con diseños de calidad en lugar de las típicas implementaciones básicas, como la inclusión de cuestionarios durante distintas etapas del juego.

También cabe destacar la necesidad de métodos de evaluación más sofisticados. Aunque los conjuntos de datos generados son amplios y ofrecen un gran potencial para aplicar técnicas innovadoras, los investigadores suelen emplear métricas e indicadores simples que no logran captar la complejidad y naturaleza contextual de ciertas habilidades y comportamientos, como las mencionadas habilidades del siglo XXI. Sin embargo, para que estos métodos avanzados funcionen adecuadamente, se requieren conjuntos de datos de tamaño considerable, y dado que muchos estudios en la literatura reportan limitaciones en el tamaño de las muestras, este sigue siendo un desafío importante en el campo. Además, muchas de estas técnicas emergentes son percibidas como “inexplicables” por los actores no técnicos, lo cual también limita su aplicabilidad en entornos reales. Por lo tanto, es fundamental seguir explorando técnicas de explicabilidad que proporcionen a estos usuarios una comprensión clara de cómo funcionan estos modelos, promoviendo la confianza, la transparencia y la toma de decisiones informadas en contextos prácticos.

Como parte del trabajo futuro, la tesis propone tres posibles líneas de trabajo para continuar esta prometedora línea de investigación: en primer lugar, el desarrollo de marcos de trabajo para el diseño e integración escalable de evaluaciones basadas

en juegos, haciendo énfasis en la alineación de las mecánicas de juego con los objetivos de aprendizaje, los métodos de evaluación y los planes de estudio existentes para facilitar su adopción en entornos educativos; en segundo lugar, la validación empírica de soluciones GBA en contextos reales, centrándose en la evaluación de la validez, la usabilidad y la accesibilidad de estas soluciones para fortalecer el marco interoperable propuesto; en tercer lugar, el avance de enfoques Human-in-the-Loop (HITL) mediante el uso de la herramienta de etiquetado desarrollada para crear modelos de evaluación más sofisticados, capaces de capturar comportamientos complejos del alumnado y apoyar el entrenamiento de sistemas de IA confiables; y por último, la integración de datos multimodales en los enfoques de evaluación actuales para enriquecer los modelos mediante evidencias provenientes de múltiples fuentes de datos.

Contents

| | | |
|----------|--|------------|
| 1 | Introduction | 25 |
| 2 | Objectives | 29 |
| 3 | Methodology | 33 |
| 4 | Results | 37 |
| 5 | Publications | 43 |
| 1 | GBA: Current Trends and Challenges | 45 |
| 2 | Towards Semantic Interoperability | 63 |
| 3 | A Framework for Interoperable GBAAA S | 81 |
| 4 | Integrating Explainable AI in Performance Prediction | 103 |
| 5 | Optimizing Manual Labeling in GBA | 119 |
| 6 | Conclusions and future directions | 129 |
| | Bibliography | 133 |

Introduction

Technology is rapidly changing the world we live in, reaching almost every aspect of our daily lives, including education. Different digital tools and platforms have been created to support these changes, including Learning Management Systems (LMSs) such as Moodle and Canvas, online learning platforms, and collaborative tools such as Google Workspace [1]. One prominent example of this digital transformation is the growing use of games for learning purposes [2]. Although the use of digital games is a relatively new trend, the idea of playing a game dates back to the ancient past and is considered an integral part of all societies [3]. Today, video games have become a widespread and influential element of everyday life, particularly among families and younger generations. In the United States, 78% of households have at least one person who plays video games, and in Europe, 53% of people aged 6 to 64 years play video games on a regular basis [4], [5].

Although games are traditionally conceived as meaningful entertainment experiences, there is ample evidence that they also serve as powerful tools for learning, skill development, and assessment [6]. Specifically, Serious Games (SGs), which do not have entertainment, enjoyment, or fun as their primary purpose, have gained significant attention over the past decade, and their versatility and adaptability make them a valuable tool across various contexts and domains. SGs have become particularly popular in educational settings, where previous research has shown that games can be more effective in learning than other traditional teaching methods due to their inherent ability to promote engagement, motivation and active participation [7]. Moreover, SGs are increasingly being adopted in corporate environments, and they play an important role in healthcare for education, prevention and rehabilitation purposes [8], [9].

SGs are currently being explored for their potential to provide assessments that are as valid as traditional methods such as standardized tests, multiple-choice exams, and self-report surveys [10]. In particular, Game-Based Assessment (GBA) is a specific application of games, referring to a type of assessment that uses players' interactions as evidence to make inferences about their knowledge and skills [11]. Games are specially attractive as assessment tools because they allow the recreation of realistic and immersive environments that closely mirror real-world contexts, supporting an appropriate application of knowledge and skills. Additionally, GBAs can embed measurement into the game mechanics, allowing for what is known as stealth assessment [12]. This approach allows for the use of unobtrusive methods (e.g., eye tracking and log files) to continuously collect data without interrupting the player's

experience [13]. Moreover, SGs are inherently engaging and motivating, and when students are actively involved and interested, the assessments they complete are more likely to reflect valid and reliable evidence of their true abilities [14].

Through GBAs, it is possible to measure a broad range of skills and constructs, including competencies identified as essential for success in real-world contexts, such as communication, teamwork and leadership, as well as 21st-century skills like creativity, critical thinking, or persistence [10], [15]. To measure such skills, conventional methods, such as standardized tests or self-report inventories, offer static snapshots that lack the granularity and real-time feedback needed for formative assessment. In contrast, SGs can generate large amounts of granular interaction data, even within a short gameplay session, which can serve as a rich source of evidence to assess what players know and are able to do [16]. Additionally, examining individual actions as well as more complex sequences of behaviors can reveal patterns of engagement and offer valuable insights into the learning process [17]. As a result, GBAs are well-suited for assessing complex, process-oriented skills that are difficult to measure using traditional approaches.

However, before the potential of GBAs can be realized, there is still a number of challenges to be addressed. The implementation of assessment features into Game-Based Learning (GBL) environments is only in its early stages because it adds a very time-consuming step to the design process [18]. Firstly, from a game design perspective, it is important that the mechanics target the emotional, behavioral, and cognitive aspects of the learner, which should be carefully assessed at different stages of the learning process [18]. Secondly, the GBA machinery (including data design, algorithms, dashboards, and other types of analytics) is usually designed specifically for each game, which increases the cost, time, and effort required for the implementation of such assessment features [19]. Although the large amounts of granular data available are promising, standardized methods for transforming this data into inferences about knowledge, skills, and attributes are not well-developed [20]. This limits the scalability and generalizability of GBAs and underscores the need for robust analytical approaches.

Usually, researchers and practitioners rely on different techniques to analyze game data. Log file analysis is often the first step, involving the implementation of basic rule-based methods that use features or metrics derived from event-level data such as clicks, movements, decisions, and time spent on tasks. More advanced techniques have emerged to capture the complexity of player interactions and competencies, such as Machine Learning (ML), Deep Learning (DL), or process mining [21], [22]. ML and DL methods can automatically detect patterns in gameplay data and classify behaviors, predict certain outcomes, or infer player characteristics. Meanwhile, process mining considers both the final outcome achieved by the student and the entire process followed to obtain it. These insights can be used not only for summative assessment purposes, but also to provide meaningful feedback to learners or even to dynamically adapt games based on individual needs [21]. However, previous research has noted that GBAs typically rely on basic analyses, and more advanced techniques such as process mining and ML remain underutilized [20].

In this Doctor of Philosophy (Ph.D.) thesis, we aim to analyze the current state of the GBA area and build upon current limitations. The first contribution involves

a comprehensive systematic review of current trends, challenges, and existing research gaps, offering a detailed overview of the field. Based on the findings of this review, we propose a set of innovative solutions to advance both the theory and practice of GBA. First, we present an interoperable semantic model that enables a consistent and structured representation of GBA data and processes. Building on that model, our work includes a scalable framework designed to support interoperability across different games assessment systems. Next, we incorporate eXplainable Artificial Intelligence (XAI) techniques to enhance the interpretability of ML models that predict student performance, making the assessment process transparent for learners and educators. Finally, we propose a practical tool to streamline the collection of manually labeled data, which aims to facilitate the implementation of ML techniques which aims to facilitate the implementation of ML techniques by significantly reducing the time and effort required to prepare high-quality training datasets. These contributions provide a comprehensive approach to addressing current challenges in the GBA field, creating new opportunities for more effective and scalable assessment practices.

This research is validated through a series of peer-reviewed articles that jointly build this Ph.D. Thesis in compilation, being the Ph.D. candidate the main author in all of them:

1. M. J. Gomez, J. A. Ruipérez-Valiente and F. J. G. Clemente, “**A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges**,” in IEEE Transactions on Learning Technologies, vol. 16, no. 4, pp. 500-515, Aug. 2023, 10.1109/TLT.2022.3226661 [23]
2. M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “**Developing and validating interoperable ontology-driven game-based assessments**,” Expert Systems with Applications, vol. 248, p. 123370, 2024. 10.1016/j.eswa.2024.123370 [24]
3. M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “**A framework to support interoperable Game-based Assessments as a Service (GBAaaS): Design, development, and use cases**,” Software: Practice and Experience, vol. 53, no. 11, pp. 2222–2240, 2023.10.1002/spe.3254 [25]
4. M. J. Gomez, Á. Armada Sánchez, M. Albaladejo-González, F. J. G. Clemente, and J. A. Ruipérez-Valiente, “**Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games**,” Expert Systems, vol. 42, no. 3, p. e70008, 2025. 10.1111/exsy.70008 [26]
5. M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “**Optimizing multimedia and gameplay data labeling: A web-based tool for Game-Based Assessment**,” SoftwareX, vol. 27, p. 101763, 2024, 10.1016/j.softx.2024.101763 [27]

Objectives

The primary motivation of this Ph.D. thesis is to explore and advance the potential of GBA as an innovative solution for evaluating knowledge, skills, and behavior in digital environments. This research aims to address the current GBA limitations regarding scalability, interoperability, and interpretability. In this sense, the dissertation aims to expand our understanding of GBA, focusing on its design considerations, practical implementations, and analytical possibilities. Through this lens, the study addresses both the conceptual and technical aspects of GBA, focusing on how game-based environments can provide valid, meaningful and adaptive assessment experiences. This work is driven by the following core research objectives:

O1. Review the current state of GBA

With the rise in popularity of games and digital learning, novel applications for assessing knowledge and skills have emerged, giving rise to GBA as a promising alternative to traditional evaluation methods. However, there was no existing study summarizing the current state of the field in this rapidly evolving area.

Therefore, *Objective 1* consists in systematically analyzing the current state of research in the GBA field, identifying key trends, challenges, and gaps in the literature that hinder its broader adoption and impact across various contexts. This thesis examines the potential benefits of GBA and proposes several ways in which it could be established as a reference model for future assessment practices. In line with this objective, the research also highlights the main limitations associated with implementing GBAs. This foundational analysis serves as the basis for the subsequent contributions presented throughout the thesis.

O2. Design an interoperable semantic model for GBA log data

One of the key open issues in the area is the specificity of GBA implementations. Usually, literature relies on custom-designed data structures and log formats tailored to individual games. This lack of standardization limits interoperability and data reuse, making it difficult to generalize findings or scale assessment solutions.

Thus, *Objective 2* aims to create an interoperable model that can contribute to the standardization of the GBA field by integrating log data from a wide variety of SGs into a single knowledge model. In addition, to ensure its effectiveness, this

objective also intends to validate the model using widely accepted metrics from the literature, demonstrating its applicability and relevance to diverse contexts and scenarios.

O3. Build a scalable framework for interoperable GBAs

The growing use of SGs has led to the creation of large data repositories, presenting a wide range of new assessment opportunities. Granular data from users' interactions with games paves the way for the application of more advanced techniques capable of detecting nuanced information related to users' cognitive skills and behavior. However, these techniques typically require large amounts of data to perform well, and data processing suffers from performance issues when the dataset exceeds the memory capacity of a single machine.

In this sense, *Objective 3* aims to develop an efficient system capable of processing large amounts of data and performing interoperable GBAs, significantly simplifying the design process. With this objective in mind, this dissertation seeks to design and implement a novel framework that leverages the previously developed semantic model to integrate data from different games into a single, efficient system capable of providing valuable insights into users' skills and behavior at scale.

O4. Enhance Predictive Models through XAI

Data from SGs can also be used to provide timely and personalized feedback to learners, or by educators to provide targeted assistance and support when needed [28]. Although Artificial Intelligence (AI) techniques such as ML have demonstrated strong predictive capabilities in this context, explainability remains an inherent challenge of these approaches. This is specially important in GBA environments, where educators and learners are often non-technical users who require clear and understandable explanations of how such predictions are made.

Motivated by this limitation, XAI emerges as a powerful tool for producing more transparent models while maintaining high performance. That way, XAI empowers non-technical users to better interpret and understand AI-generated insights, enabling them to make informed decisions [29]. Thus, *Objective 4* seeks to integrate XAI techniques into GBA environments to enhance the interpretability of learner performance predictions. Through this contribution, the thesis aims to provide a comprehensive framework for interpretable models in GBAs, supporting a better understanding of AI-driven assessments and facilitating informed decision-making.

O5. Optimize data labeling for AI techniques in GBA

GBAs can collect various types of data, including audio recordings, video captures, and log data from player interaction with games. As previously discussed, AI models offer stakeholders a wide range of opportunities for learner assessment. In this context, labeled data plays a crucial role in the development of such AI models and algorithms. However, researchers usually employ primitive methods, such as Excel worksheets or manual annotations, as the data labeling process is time-consuming and can be very costly, specially when expert annotators are required

[30]. Although previous approaches for optimizing the labeling process have been explored, existing data labeling tools are often too generic to address the specific needs of GBA scenarios.

To address this gap, *Objective 5* focuses on the design and development of a practical tool adapted to the specific data requirements of GBA scenarios, with the goal of enhancing the efficiency and accuracy of the data labeling process.

By completing these objectives, this research will contribute significantly to the scholarly discourse on assessments using digital games, paving the way for more scalable, interoperable, and interpretable GBA solutions that can be effectively integrated into educational, training, and other applied contexts.

Methodology

This chapter introduces the methodology followed in this Ph.D. thesis. The methodology was conducted following a scientific approach based on the continuous study of the state of the art and the analysis of the results obtained during the different stages of the research. This thesis is defined as a set of five papers published in high-impact journal indexed in the Journal Citation Reports (JCR).

M1. Systematic literature review

The first step in addressing [*Objective O1*] of this thesis was to provide a comprehensive overview of the research field. To achieve this, a systematic review of the existing literature was conducted, aiming to offer a detailed understanding of current trends, key challenges, and potential future directions in the field.

The review conducted a detailed analysis of recent research in the GBA area, following a standard systematic literature review methodology based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [31]. After formulating a set of Research Questions (RQs), we applied a predefined set of search queries across several identified bibliographic databases, along with clear inclusion and exclusion criteria. This analysis covered a broad range of academic publications, including journal articles, book chapters, and conference proceedings, ensuring a wide coverage of the topic. The collected literature was comprehensively categorized and analyzed based on specific criteria, such as the context in which the GBA was applied, the primary purpose of the assessment, and the methods or techniques employed. This granular categorization facilitated the identification of prevalent trends, frequently used approaches, and potential gaps within the current body of knowledge. These gaps form the basis for the subsequent studies presented in this dissertation.

This step in the methodology highlighted open challenges in GBA research and concluded with key insights and future research directions. The findings are validated through the following publication, available in Section 5.1:

- M. J. Gomez, J. A. Ruipérez-Valiente and F. J. G. Clemente, “**A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges**,” in IEEE Transactions on Learning Technologies, vol. 16, no. 4, pp. 500-515, Aug. 2023, 10.1109/TLT.2022.3226661 [23]

M2. Semantic model design and implementation

A crucial aspect of GBA is the nature and structure of the data used, as both the type and format of the data directly affect the validity, interpretability, and outcomes of the assessment. One key finding from the systematic review was that very few studies reported details about the type or format of the data employed. This is often due to the fact that researchers use specific data formats and customized methods tailored to individual games, making the integration of assessment into games both costly and time-consuming.

In response to this gap, and to address *Objective O2*, we designed and implemented an ontology-driven semantic model aimed at standardizing log data formats in GBA. This model enabled the integration of data from a variety of games into a unified ontology framework. For the development of the ontology, we adopted Methontology [32], a structured methodology recognized as the most mature approach for building ontologies and recommended by the Foundation for Intelligent Physical Agents (FIPA). Following the creation of the ontology, we conducted a formal evaluation to identify potential design and technical issues in the model.

In addition, to validate the model's suitability and applicability, extensive simulations and real-world tests were conducted, evaluating its functionality across diverse contexts and with various data formats. This validation study incorporated a selection of previous metrics from literature, as well as the design and implementation of novel metrics that demonstrated the assessment and generalization capabilities of the presented approach. The results are validated in the following publication, available in Section 5.2:

- M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “Developing and validating interoperable ontology-driven game-based assessments,” Expert Systems with Applications, vol. 248, p. 123370, 2024. 10.1016/j.eswa.2024.123370 [24]

M3. Interoperable framework development

The systematic review revealed that most assessments are conducted with small sample sizes, resulting in low statistical power and limited generalizability. To support more meaningful and generalizable assessments, it was essential to provide a scalable and interoperable framework that allows the application of assessment techniques without concerns about performance or scalability constraints. Given the developed ontology-driven model, the immediate next step was to leverage this semantic model to design an interoperable framework capable of aggregating and analyzing data across different games.

This step addresses *Objective O3*, and involved the creation of a framework to fulfill five key requirements: 1) the integration of an intermediate semantic layer; 2) the ability to process large-scale data; 3) interoperability for GBA metrics and visualizations; 4) easy communication with external sources; and 5) support for privacy, authentication, and authorization configurations. To integrate data from different games into an ontology-compatible format, a preprocessing stage was performed in which the data is transformed into RDF/XML format. The developed architecture

incorporates the existing SANSA framework as a basis for performing our analysis. SANSA [33] is an open-source structured data processing engine that enables distributed computation over large-scale RDF datasets. Within our architecture, we leveraged SANSA’s capabilities to perform interoperable assessments; for example, using the *inference library* to infer new information from the existing RDF data, or the *querying library*, which provides methods for performing queries directly over the constructed RDF graphs.

Additionally, this step involved a case study validation in which the output from the architecture was used to build a learner report system and exemplify how the assessment architecture can be used in real contexts. The results are published in the following article, available in Section 5.3:

- M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “**A framework to support interoperable Game-based Assessments as a Service (GBAaaS): Design, development, and use cases**,” *Software: Practice and Experience*, vol. 53, no. 11, pp. 2222–2240, 2023.10.1002/spe.3254 [25]

M4. Interpretability in predictive models

One of the remaining challenges in the field was the development of new approaches to improve interpretability in GBA environments. In this regard, this step of the methodology focused on enhancing the interpretability of ML models used to predict learner performance. The specific methodology was structured into two main components: performance prediction model development and model interpretability and explanations.

The first component refers to the methodology followed to build the ML models. A set of features was designed to predict users’ performance in a SG, specifically whether a user would successfully complete the level being played. To achieve this, a multi-prediction approach was employed, based on distinct time intervals derived from the average completion time of each level. These intervals were defined at 25%, 50%, and 75% of the average completion time, enabling the real-time monitoring of player progress at different points during gameplay. Based on these intervals, a set of features was created and categorized into three groups: *user features*, which provide information about the user’s overall performance; *level features*, which capture the unique characteristics of each level; and *attempt features*, which provide detailed information about the user’s actions and decisions during a specific level attempt. Then, we considered a set of ML algorithms for training and selected the models and configurations associated with the algorithms that achieved the best average results.

The second component refers to the methodology followed to achieve model interpretability. If the best model was inherently interpretable, its interpretability was leveraged to explain the model’s predictions. However, if the selected model lacked inherent interpretability, XAI techniques were applied to generate meaningful explanations for its predictions. In particular, a post hoc, model-agnostic, and locally interpretable method was employed. This approach allowed for the interpretation of different models (model-agnostic) by analyzing their behavior on individual predictions (locally interpretable) after the model has been trained (post hoc). Specifically,

we selected SHAP, a method for explaining individual predictions based on Shapley values from cooperative game theory, which quantifies the contribution of each feature to a particular prediction in a consistent and theoretically grounded manner [34].

The results of this study address [*Objective O4*] and are presented in the following publication (see Section 5.4):

- M. J. Gomez, Á. Armada Sánchez, M. Albaladejo-González, F. J. G. Clemente, and J. A. Ruipérez-Valiente, “**Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games,**” *Expert Systems*, vol. 42, no. 3, p. e70008, 2025. 10.1111/exsy.70008 [26]

M5. Implementation of an open-source practical tool to enhance the labeling process

In the context of GBA, human-labeled data is often underutilized, limiting the effective application of AI techniques. To address *Objective O5*, an open-source tool was developed to optimize the labeling process in this domain. This involved the creation of a web-based application specifically designed for annotating GBA data, with support for *audio*, *video*, and *game-event* streams.

The tool is presented comprehensively in the following article (see Section 5.5):

- M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “**Optimizing multimedia and gameplay data labeling: A web-based tool for Game-Based Assessment,**” *SoftwareX*, vol. 27, p. 101763, 2024, 10.1016/j.softx.2024.101763 [27]

M6. Conclusions and future work

This PhD thesis concluded with a synthesis of the main findings and a discussion of their implications for GBA research and practice. The conclusion also highlighted potential directions for future work, particularly in the context of rapidly evolving educational technologies and the increasing availability of gameplay data.

Results

In this chapter, the key findings and major contributions of this Ph.D. are carefully outlined and discussed.

R1. The current state of GBA

The first publication of this thesis (A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges) presents a systematic review of the current state of the field (*Step M1*), based on an analysis of 65 research papers published between 2013 and 2020. This study first identifies that GBAs are applied in different contexts, with the most common being *K-16 education* (specifically high school and middle school), the *workplace*, and *medical settings*. Regarding the purpose of the assessment, the review finds that most studies focused on using games to evaluate learning outcomes, while also aiming to demonstrate that these assessments represented valid measures within real educational environments.

The analysis also reveals four major domain categories across the reviewed research papers: *STEM*, *humanities and social sciences*, *cognitive and soft skills*, and *physiological capacities*. The survey further revealed that many of the studies had small data samples, with nearly half of them explicitly reporting data sample limitations. Directly influenced by the previous finding, the review found that most studies (80%) employed *descriptive statistics* as their primary analytical method, while a smaller proportion (36%) utilized ML techniques, and only a single study applied DL methods.

In addition, the discussion highlights several open challenges in the field based on the reported findings. One of the most significant challenges is related to replication and the difficulty of transferring research into practice. The systematic review revealed that, in most cases, the game or tool used in the studies was not made publicly available, and critical information about the datasets was often missing. Furthermore, many studies lacked transparency in key methodological aspects. Regarding the applied methods, the review underscores the need to explore more advanced analytical techniques. However, this is constrained in current research due to the limited size of data samples used. Moreover, as researchers increasingly adopt such techniques, the survey also emphasizes the growing need for more interpretable models to ensure transparency and facilitate understanding of the results.

In conclusion, the systematic review effectively maps out the current landscape of assessments using digital games, identifies the methodologies employed to analyze

collected data, and highlights both the potential and challenges of leveraging game-based environments for modeling user behavior and assessing learning progression.

R2. Interoperable ontology-driven GBAs

The lack of information about datasets and key methodological aspects underscores the need for an interoperable framework to conduct GBAs. As a result of step *M2*, the second publication of this Ph.D. thesis (Developing and Validating Interoperable Ontology-driven Game-Based Assessments) presents a novel ontology that conceptualizes the core concepts and relationships of the GBA domain, such as *Game*, *Scenario*, *User*, and *UserGroup*. Moreover, relationships between concepts, such as that between *User* and *UserGroup*, are designed to enhance the model and can be further leveraged to perform ontology queries involving reasoning tasks.

To practically validate the ontology, we used two groups of metrics: those derived from previous literature, such as *activity indicators*, *event types*, and *user performance*, and newly designed metrics, including *persistence* and *play styles*. The initial calculations required for these metrics are implemented in form of SparQL (the standard language for querying RDF data) queries. More advanced techniques, such as ML, can also be integrated into the system by developing separated scripts that leverage the results obtained from the SparQL queries.

Finally, to illustrate how this ontology-based approach can be applied in a real environment, this research presents a case study that tests its interoperability and usability using data from ten different SGs, integrated within a visualization dashboard system. This dashboard allows instructors to monitor learners' activities while playing, use the collected data to adapt their interventions when needed, or even incorporate the resulting metrics into formative assessments.

Notably, this research is pioneering in enabling interoperable assessments using data from diverse SGs, through the development of an ontology-based semantic model that allows the integration of game data into a unified framework. Compared to the limitations identified in the systematic review, this approach paves the way for greater interoperability, reusability, and reproducibility in GBA research and practice.

R3. A framework to support scalable and interoperable GBAs

Following the completion of the previous study, step *M3* presents a novel approach that combines the use of ontologies with a big data architecture to perform interoperable GBAs. The third core article of this dissertation, A framework to support interoperable Game-based Assessments as a Service (GBAaaS): Design, development, and use cases, presents two main contributions. First, it details the development of a scalable and interoperable framework that uses SANSStack and the previously developed ontology as a baseline. Second, it describes the evaluation of the framework, focusing on both system performance and usefulness through a case study.

The proposed framework has five main components. First, the *preprocessing module* transforms the raw data, which can be received in various standard formats such as CSV, TSV, or JSON, into RDF format, making it compatible with the ontology-based knowledge model. Once the data is converted into the appropriate format, the *analytics, inference, and querying module* reads the data in the form of RDF triples in order to process it as a distributed data structure. To enrich the existing knowledge in the dataset, the system allows the definition of custom rules, which can be used to infer new knowledge from existing facts. After the inference is completed, the *metrics module* computes the defined in-game metrics using SparQL queries. The results can then be exported in a human-readable text format, saved as CSV files, or stored in a MySQL database.

The fourth component is the *authentication and authorization module*. This module enables access control to specific resources based on user roles. Specifically, the system defines three distinct roles: the admin role, which has full access to system functionalities; the instructor role, which can insert new GBA data and query metrics related to the games and groups they are involved in; and the learner role, which is limited to querying their own metric results. Finally, the system includes a *Service API* that allows it to be accessed as an external service, introducing the paradigm of Game-Based Assessment as a Service (GBAaaS). The API provides methods for retrieving metric data and for submitting new data to be processed by the system. For data submission, two types of endpoints are available: one designed for uploading complete datasets and another one for integration with streaming-oriented systems (e.g., a user is playing a game and the system sends individual events to the API in real time).

The performance evaluation was conducted using different cluster configurations and dataset sizes to assess how the architecture was able to handle incoming data. The results indicated that the cluster of one master node and four worker nodes offered the best balance between resource management and performance. This configuration was capable of processing two million user events (equivalent to data generated by 39 classrooms using a game for one hour per week over the course of a month) in an average of 107.2 minutes. In addition, the case study validation included the implementation of a dashboard and a reporting system, both of which consumed the *Service API*. These two components showed the benefits of interoperability between games and metrics, as well as the effectiveness of the role-based access control implemented in the architecture, which ensures that each user only sees the information relevant to their role.

R4. Interpretability in real-time predictive models

Derived from methodological step *M4*, the fourth publication (Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games) introduces a novel methodology for improving interpretability in real-time performance prediction models. The configuration and evaluation of the AI algorithms, conducted through a ten-fold cross-validation, demonstrated that the Random Forest (RF) algorithm achieved the best performance across all evaluated time intervals (25, 50 and 75% of the average level completion time). Specifically, the

RF model obtained a balanced accuracy of 0.76 at the 25% interval, 0.772 at 50%, and 0.795 at 75%. It is noticeable that the models exhibited good performance even in the earliest time interval, when only limited information was available. The test results were consistent with the validation outcomes, with the third time interval model achieving the highest balanced accuracy (0.793).

Since RF is considered a “black box” algorithm, the study employs the SHAP method to explain individual predictions. This approach enables the identification of the most relevant factors contributing to students’ success in solving different levels. The results revealed that the most critical features across all predictions were related to the level difficulty and the number of actions performed by the user during gameplay. This means that the model’s prediction is mainly influenced by these features, although other features, such as the number of previously completed levels and the time spent on the current level, also contribute to the prediction.

Finally, the research presents a use case to illustrate how the enhanced interpretability can be applied in a classroom setting to support individual students. This is achieved by providing the teacher with a visualization that shows the model’s current prediction along with the features contributing to that prediction. In this way, if the teacher observes that a student is struggling with a specific puzzle for identifiable reasons, they have the opportunity to offer personalized support if necessary.

Results from this research demonstrate that the developed ML models can anticipate students’ task completion in a SG, making accurate early predictions even after a short period of gameplay. Moreover, the study highlights how these predictions can be made fully explainable by incorporating both intrinsic and extrinsic explainability options. In doing so, this work offers a framework for interpretable models in this domain, enabling a deeper understanding of AI model predictions and supporting informed decision-making in educational contexts.

R5. GBA Labeling Tool

As a result of step *M5*, the fifth publication of this Ph.D. thesis (Optimizing multimedia and gameplay data labeling: A web-based tool for Game-Based Assessment) introduces a novel open-source labeling tool specifically designed for annotating various types of GBA data. The tool is developed using the *Django* framework and employs a SQLite embedded database, which is well-suited for most low to medium traffic applications. The application supports *audio*, *video*, and *game-event* data, and uses a custom parser to transform the raw game-event data into structured *event* instances within the database. In addition, the tool includes a *feature computing* module that analyzes the data and automatically calculates a set of relevant features to provide context and support during the labeling process.

Since the way data is visualized during annotation is crucial, the application provides multiple visualization options. Specifically, *game-event* data can be visualized in three different ways. First, the tool includes templates that support integration with *Unity WebGL applications*, allowing the use of the game engine itself to replay and visualize the gameplay. This offers annotators a rich, interactive way to observe player actions as they occurred in-game. Second, the tool can generate a

textual (“pretty-printed”) representation of the game events. This representation includes key information such as the name of each action and its timestamp relative to the previous action. Moreover, these *text replays* are fully customizable, since annotators can define new event types by combining existing ones along with a set of regular expression operators, enabling flexible and specific representations of gameplay sequences. Finally, the third visualization method involves plotting the game-event data. These plots offer a graphical representation of the sequence of actions over time, helping annotators to quickly locate patterns and key moments in the data.

Regarding annotations themselves, we define three types of annotations: *global annotations*, which indicate that a specific label applies to the entire replay; *time-instant annotations*, which indicate the occurrence of a specific label at a particular moment within the gameplay; and *time window annotations*, which refer to labels that span a defined time interval between beginning and end of the replay. These annotation types allow for flexible labeling, accommodating a wide range of use cases and research objectives. Finally, the tool provides functionality to export the annotated data in both JSON and CSV formats, allowing researchers to easily integrate the data with external analysis pipelines.

Publications

1 GBA: Current Trends and Challenges

Title

A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges

Authors

Manuel J. Gomez¹, José A. Ruipérez-Valiente¹,
Félix J. García Clemente¹

¹*Department of Information and Communications Engineering,
University of Murcia, Spain*

Publication details

| | | | |
|----------------|--|------------------|--------------------------|
| Journal | IEEE Transactions on Learning Technologies | Publisher | IEEE |
| Volume | 16 | Number | 4 |
| Pages | 500-515 | Year | 2023 |
| JIF | 2.9 | Rank | Q2 |
| Status | Published | DOI | 10.1109/TLT.2022.3226661 |

Abstract

Technology has become an essential part of our everyday life, and its use in educational environments keeps growing. In addition, games are one of the most popular activities across cultures and ages, and there is ample evidence that supports the benefits of using games for assessment. This field is commonly known as game-based assessment (GBA), which refers to the use of games to assess learners' competencies, skills, or knowledge. In this article, we analyze the current status of the GBA field by performing the first systematic literature review on empirical GBA studies. It is based on 65 research papers that used digital GBAs to determine: the context where the study has been applied, the primary purpose, the domain of the game used, game/tool availability, the size of the data sample, the computational methods and algorithms applied, the targeted stakeholders of the study, and what limitations and challenges are reported by authors. Based on the categories established and our analysis, the findings suggest that GBAs are mainly used in K-16 education and for assessment purposes, and that most GBAs focus on assessing STEM content, and cognitive and soft skills. Furthermore, the current limitations indicate that future GBA research would benefit from the use of bigger data samples and more specialized algorithms. Based on our results, we discuss current trends in the field and open challenges (including replication and validation problems), providing recommendations for the future research agenda of the GBA field.



A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges

Manuel J. Gomez [✉], José A. Ruipérez-Valiente [✉], *Senior Member, IEEE*, and Félix J. García Clemente [✉]

Abstract—Technology has become an essential part of our everyday life, and its use in educational environments keeps growing. In addition, games are one of the most popular activities across cultures and ages, and there is ample evidence that supports the benefits of using games for assessment. This field is commonly known as game-based assessment (GBA), which refers to the use of games to assess learners' competencies, skills, or knowledge. In this article, we analyze the current status of the GBA field by performing the first systematic literature review on empirical GBA studies. It is based on 65 research papers that used digital GBAs to determine: the context where the study has been applied, the primary purpose, the domain of the game used, game/tool availability, the size of the data sample, the computational methods and algorithms applied, the targeted stakeholders of the study, and what limitations and challenges are reported by authors. Based on the categories established and our analysis, the findings suggest that GBAs are mainly used in K-16 education and for assessment purposes, and that most GBAs focus on assessing STEM content, and cognitive and soft skills. Furthermore, the current limitations indicate that future GBA research would benefit from the use of bigger data samples and more specialized algorithms. Based on our results, we discuss current trends in the field and open challenges (including replication and validation problems), providing recommendations for the future research agenda of the GBA field.

Index Terms—Educational technology, game-based assessment (GBA), game-based learning (GBL).

I. INTRODUCTION

TECHNOLOGY is progressively changing the world in which we live. Over the last decade, it has started to make a significant impact on educational environments, and increasing evidence has been accumulated showing the positive impact of technology in education [1]. One of the most prominent examples of technology in education is the use of digital games [2]. This type of games has become a significant part of families and, especially, among young people around the world. In fact, three-quarters of all U.S. households have at least one person who plays video games [3], while in Europe, 51% of the population aged 6–61 years play video games (an average of 8.6 h/week) [4]. Moreover, many educators see digital games as

powerfully motivating digital environments because of their potential to enhance student engagement and motivation in learning [5]. This increasing interest provides an opportunity to use video games as a tool to improve learning and education. Specifically, there is much enthusiasm in the field of education about game-based assessment (GBA) because conventional assessment methods do not seem to fully have the power to measure all aspects of students' knowledge, skills, and attributes [6].

Accompanying this explosion in technology use is the quantity, range, and scale of data that can be collected, which have increased exponentially over the last decade [7]. In education, the increase in e-learning resources, educational software such as Google Classroom or Kahoot, and the use of the Internet have created large repositories that provide a goldmine of educational data that can be explored and used to understand how students learn [8]. Regarding games, they allow recreating more authentic situations compared to traditional classroom activities, such as lectures or written exercises. From these situations, we can collect a vast amount of detailed data on students' interaction with the game, which provides a great opportunity to make GBAs in ways that are not possible in traditional testing [9].

Over the last ten years, numerous studies (see the work in [10] for a meta-analysis) have reported that games can be more effective for learning than other traditional teaching methods. In addition, when measuring the competencies acquired, most traditional tests present individual and decontextualized items to learners, while 21st-century competencies benefit from being applied in context for more accurate measurements. Furthermore, classic assessment often interrupts the learning process, and it does little to motivate learners [11]. Since digital games often employ challenging, interesting, and complex problems, they can be used to generate evidence of 21st-century competencies, which are traditionally difficult to measure using conventional forms of assessment [12]. The advantages of using games as a form of assessment are manifold [11], [13], [14]: they are engaging and motivating (which provides more valid assessments), and they allow us to create more complex and authentic scenarios required to assess the application of knowledge and skills. Moreover, immediate feedback based on learners' activity can reveal teachers' specific areas of difficulty to make learners keep up with the pace of the class, and such assessment would result in an adaptive game environment, which changes with learners' activity.

The implementation of assessment features into game environments is only in its early stages because it adds a very time-consuming step to the design process [15]. This situation

Manuscript received 10 December 2021; revised 26 September 2022 and 30 November 2022; accepted 1 December 2022. Date of publication 5 December 2022; date of current version 16 August 2023. (Corresponding author: Manuel J. Gomez.)

The authors are with the Faculty of Computer Science, University of Murcia, 30100 Murcia, Spain (e-mail: manueljesus.gomez@um.es; jruipere@mit.edu; fgarcia@um.es).

Digital Object Identifier 10.1109/TLT.2022.3226661

1939-1382 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See <https://www.ieee.org/publications/rights/index.html> for more information.

Authorized licensed use limited to: MIT. Downloaded on March 26, 2025 at 15:40:47 UTC from IEEE Xplore. Restrictions apply.

calls for a review of the current state of the art in the GBA field for effective implementations. In this respect, we found some previous works that performed metareviews of the existing research on the different applications of games in learning and education. For example, Qian and Clark [16] reviewed 137 papers to determine what empirical evidence existed concerning the effects of game-based learning (GBL) on 21st-century competencies and identified successful game-design elements that aligned well with established learning theories. Moreover, Alonso-Fernandez et al. [17] carried out a review focused on data science applications to game learning analytics data, showing that the primary purpose when analyzing data from serious games was assessment. Furthermore, Gris and Bengtson [18] aimed to answer how learning, engagement, and usability of games are evaluated in GBL research. To this aim, they conducted a systematic review of 91 empirical studies and categorized their measures and instruments. The researchers concluded that future research in GBL studies should add direct assessments and indirect measures to assess engagement and usability. Guan et al. [19] provided a systematic review of 35 experimental studies that substantially integrated gaming elements in primary school lessons and they noted that gamification was the most frequently used game genre. Finally, Chen et al. [20] conducted a systematic review of 146 articles related to GBL in science and mathematics education. These researchers concluded that GBL is mainly used to increase learner motivation and engagement and reduce learning anxiety. They also revealed that analyzing higher order thinking skills (e.g., problem-solving, group collaboration) is one of the main hot topics in the community.

Despite the previous reviews of the use of games in learning in education, we have not found any specific study reviewing literature about GBA. For this reason, this article aims to conduct the first systematic literature review on the applications of empirical GBA studies and answer some research questions (RQs) based on the analysis performed to discover current trends and open challenges in this area. The results obtained will provide an overall view of the GBA field, defining its current status and potential future steps in the research in this area.

The rest of this article is organized as follows. Section II describes the methods, including some terminology clarifications, the RQs, databases and search terms, research selection, as well as review process. Section III presents the analysis and synthesis of our results. Then, we end the article with a discussion in Section IV and Finally, Section V concludes this article.

II. METHODS

We followed a standard systematic literature review methodology, using the preferred reporting items for systematic reviews and meta-analyses (PRISMA) [21] as a basis for conducting our systematic review. First of all, we formulated each RQ. Then, we used a fixed set of queries on a preidentified bibliographical database, and a set of inclusion and exclusion criteria. Next, we made a full paper review and coding process of the RQs, and finally, we carried out a synthesis and analysis. No time restrictions were set. We can see a flow

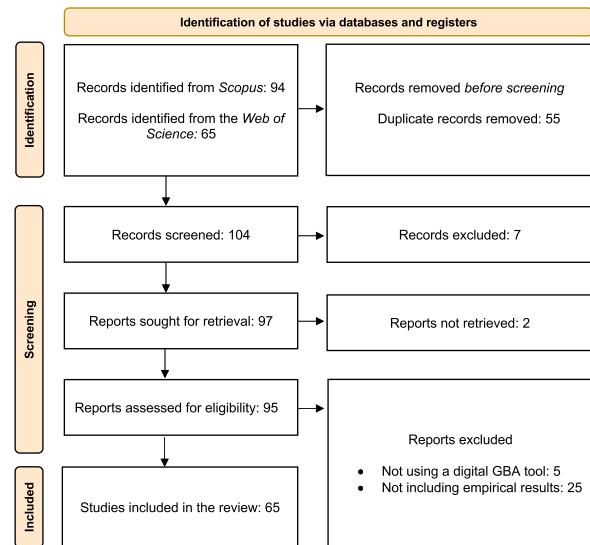


Fig. 1. Flow diagram representing the different phases of the systematic review.

diagram representing the different stages of our systematic review (following the PRISMA template [22]) in Fig. 1.

A. Terminology Clarifications

In this section, we present a set of definitions that aim to clarify the concept of GBA, which is the focus of this systematic literature review. First, we can define a *game* as “a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome” [23, p. 80]. In addition, games also have clearly defined goals and obstacles for the player to overcome, providing only intrinsic rewards (satisfaction for getting the right answer) [24]. Second, *GBL* can be regarded as an innovative learning approach where a game is developed to deliver immersive and attractive learning experiences aiming at particular learning goals, experiences, and results [25]. Thus, GBL uses a game containing learning content derived from school curricula or essential life skills to improve the learning experience. Moreover, *GBA* is a specific application of games, referring to a type of assessment that uses players’ interactions with the game, both digital and nondigital, as a source of evidence to make meaningful inferences about what players know and can do (i.e., knowledge, skills), and how individual players interact with the game as a problem-solving process [15], [26]. Finally, we have the concept of *gamification*, which is usually defined as “the use of game design elements in nongame contexts” [27, p. 9].

Although GBL and GBA are often confused with gamification and gamified assessment, it is undeniable that some differences exist between them. While GBL implies the use of a game developed for learning purposes, gamification utilizes game elements in nongame contexts, not necessarily using full games inside the activities [28]. Thus, GBA also implies the use of a game developed for assessment purposes, using players’ interaction with the game as a way to obtain evidence and

use this evidence as a form of assessment. Therefore, tools that use gamified activities to assess students' knowledge (e.g., Kahoot, Duolingo) use gamified assessments and cannot be considered as GBAs. We can also make a clear distinction between GBA and a simple measurement using games since GBA is intended for evaluating players' skills or knowledge based on their interaction with the game. As Ghergulescu and Muntean [29, p. 357] state, "measurement represents the process of collecting the information needed for assessment." In other words, measurements are used as evidence to make meaningful inferences about what players know and can do, while measurements using games do not perform that evaluation. These are the definitions that we applied as part of the systematic review screening process to consider a given paper within the GBA field or not, including or discarding that study.

B. Research Questions

To state each one of the RQs, we analyzed and simplified the steps in empirical GBA research [30], which can be seen in Fig. 2.

In this process, we can identify the following five different stages:

- 1) learning environments, with the context and learners;
- 2) the GBA tool that is going to be used in the research;
- 3) data collection, to identify which data has to be collected and how to store them;
- 4) modeling and assessment machinery;
- 5) educational application, to identify the final objective and target users.

From these stages, we identified the following RQs, which allow us to understand the open challenges and current trends in the area.

RQ1. In what context or environment has GBA been applied?

RQ2. What is the primary purpose of GBA?

RQ3. What is the domain of GBA?

RQ4. Is the game/tool used available to the public?

RQ5. What is the size of the data sample used in the study?

RQ6. What computational methods and algorithms have been applied in the research?

RQ7. What stakeholder is the intended recipient of the research results?

RQ8. What limitations and challenges do the authors address?

In addition, Fig. 2 also shows the mapping between the different stages and the RQs identified. RQ1 is based on the first stage, related to the learning context. Then, RQ2, RQ3, and RQ4 are based on the second stage, which refers to the GBA tool used, its primary purpose, the domain, and availability. Next, we wanted to investigate the sample size (RQ5) which

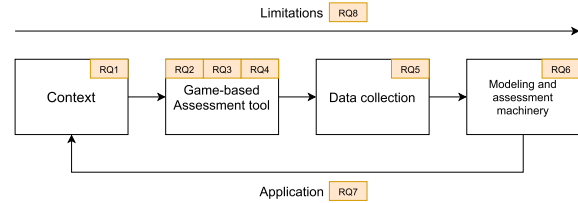


Fig. 2. Simplified view of the steps in GBA empirical research.

is situated in the data collection stage. Regarding the modeling and assessment machinery, our objective was to investigate the computational methods and algorithms applied in the research (RQ6). RQ7 refers to the application of the research in the desired context, identifying the main stakeholder of the research. Finally, RQ8 aims to identify the research limitations at any stage.

C. Databases and Search Terms

We have queried two databases: Scopus and the Web of Science since they are the most widely used databases in different scientific fields and are often used for surveying the literature [31]. Scopus is the world's largest citation database of peer-reviewed research literature, with over 22 000 titles (including journals, conferences, and book series) from more than 5000 international publishers, of which 20 000 are peer-reviewed journals in the scientific, technical, medical, and social sciences [32]. Moreover, the Web of Science, the second biggest bibliographic database, can be used to track ideas going back several decades from almost 1.9 billion cited references from over 171 million records [33].

To perform the search on both databases, we restricted the query to title and keywords: 1) we included the term "game-based assessment" and searched for it within the paper titles; 2) we included the term "game-based assessment" and searched for it within the paper keywords. Thus, we used the following final search query:

(TITLE("game-based assessment") OR KEY("game-based assessment")).

The initial selection of studies was retrieved in January 2021, and this query generated 159 initial studies (94 from Scopus and 65 from the Web of Science).

D. Inclusion/Exclusion Criteria

After obtaining the initial collection, we excluded the duplicated records from the two databases (55 studies). Then, we made a first brief review of all papers, comparing them against the inclusion and exclusion criteria. This first review was conducted by one of the authors. After the first analysis, we classified studies as *included* or *excluded*, and the coding results were discussed collaboratively by the three authors in order to obtain the final set of included and excluded studies and avoid possible errors. The inclusion/exclusion criteria followed are described as follows. Given these criteria, the paper was included if all of the following conditions were met (i.e., if one condition was not met, the paper was excluded). Furthermore, the conditions were

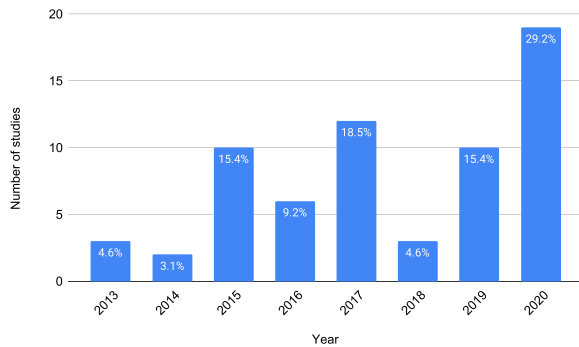


Fig. 3. Number of selected studies and rate in the collection per year of publication.

applied sequentially so that a paper not matching a condition was excluded immediately from the collection.

- 1) The paper was written in English or Spanish (languages in which the authors have high proficiency): 0% of the papers were excluded.
- 2) The paper was fully accessible: 1.9% of the papers were excluded (2 studies).
- 3) The paper was published in conference proceedings, journals, or edited books/volumes (i.e., book chapter): 0% of the papers were excluded.
- 4) The paper was not extended at a later time (i.e., a conference paper that was later on extended in a journal paper): 6.9% of the papers were excluded (7 studies).
- 5) The paper used a digital GBA tool: 5.3% of the papers were excluded (5 studies). See Section II-A for relevant definitions.
- 6) The paper included empirical evidence related to the outcomes of applying the GBA tool: 27.8% of the papers were excluded (25 studies).

E. Final Paper Collection

After the first brief review to ensure that every paper met our inclusion/exclusion criteria, we excluded a total of 39 papers. Thus, the final paper collection consists of 65 studies.

Fig. 3 shows the distribution of papers within the final collection by publication year. We see an increasing interest in this particular topic: between the years 2013 and 2016, we only have 21 (32.3%) published papers that matched our criteria, while between 2017 and 2020, there are 44 (67.7%) of them.

We also collected each paper's keywords and made a brief analysis to describe our paper collection. For our analysis, we excluded the "game-based assessment" keyword since it was the most common one. The most frequent keywords are presented in Fig. 4. The total sum of keywords is 299 while there are 201 unique keywords. The average keyword was found 1.48 times. As we can see, the predominant keywords strongly focused on games, assessment, and analytics.

F. Review and Coding Process

In the coding stage, we collected the data of the selected studies that we consider to be the most valuable to address the

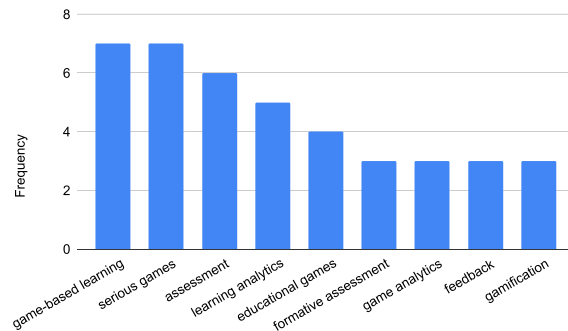


Fig. 4. Distribution of keywords across articles in the final collection.

RQs in Section II-B. Based on the aim of the review, we followed an inductive coding scheme (also called open coding). This means that the codes created were based on the qualitative data itself [34]. This is an iterative process since researchers can add new codes, split an existing code into two, or compress two existing codes into one as they continue reviewing data. Specifically, in our analysis, we first made a brief review of each paper (conducted by one author), collecting all the necessary information to code each RQ at once. After that, we followed an iterative process whereby we continued reviewing the information corresponding to each RQ sequentially, and unclear results were discussed and contrasted by the three authors. The full results of the coding process per paper are available in [35]. In addition, it should be noted that each paper can fit into more than one of the codes created for each RQ.

III. RESULTS

A. In What Context or Environment Has the GBA Been Applied? (RQ1)

GBAs can be used in very different environments. Our analysis reveals that there are three main contexts where GBAs have been used.

- 1) **K-16 education:** Some papers use GBAs in K-16 education (e.g., school, university) to support teaching and learning. More specifically, games are most commonly used in middle school and high school (23.1%). However, games are also used in other K-16 education environments such as primary school (15.4%) and university (10.8%). For example, Di Cerbo et al. [36] used game data from 751 US middle school players.
- 2) **Medical:** Games can also be used in medical environments for different purposes (e.g., rehabilitation). For example, Lindenmayer et al. [37] examined the feasibility of administering the GBA in a sample of inpatients with chronic schizophrenia with low levels of functioning. Moreover, Wiloth et al. [38] aimed to present data on construct validity, test-retest reliability, and feasibility, measuring motor-cognitive functions in multimorbid patients with mild-to-moderate dementia. Regarding construct validity, the authors tested eight hypotheses and confirmed seven of them (87.5%), thus indicating

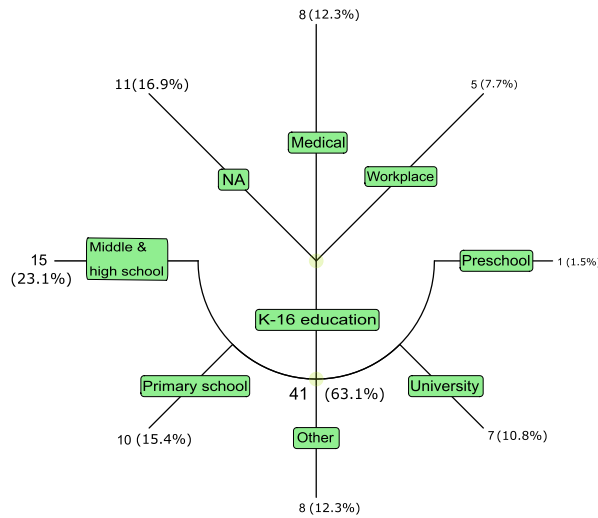


Fig. 5. Papers' category distribution based on RQ1.

excellent construct validity. Moreover, the authors found moderate-to-high test-retest reliability ($ICC=0.47-0.92$).

- 3) **Workforce:** Another option is the use of games to assess in professional environments. In this context, enterprises can use games to evaluate their employees or provide them with additional feedback. Even now, companies can include the use of GBA for the recruitment of staff and the selection process [39]. This idea is supported by the fact that in-game constructs show similar relationships with in-game performance to what the workforce constructs do with job performance [40].

There are also some studies such as [41] that did not specify in what context their games were used (11 studies, 16.9%). We can see the number of papers fitting in each category in Fig. 5. As the figure shows, GBAs are mostly used in K-16 education (41 studies, 63.1%), followed by medical (8 studies, 12.3%) and workforce contexts (5 studies, 7.7%).

B. What is the Primary Purpose of the GBA Research? (RQ2)

In this RQ, we wanted to know what was the main purpose of each GBA in each study. We coded the papers' main purpose into six different categories: GBA evaluation, study of in-game behaviors, assessment, interventions, framework proposal, and game design proposal. Next, we describe in detail each one of these categories.

- 1) **GBA evaluation:** In these studies, authors evaluate the game by checking if it achieves its initial objectives using some measure to prove that the game or tool is suitable for an educational environment. Hummel et al. [42] authors showed how they applied the methodology for an assessment game for ICT managers in secondary vocational education, checking if this assessment was content-valid compared to a face-to-face assessment. Moreover, Marengo and Pagano [43] aimed to investigate whether it is possible to perform an in-Basket test (which is widely used by companies and organizations

in order to map employees' soft skills) online with the same effect as that of the onsite one.

- 2) **In-game behaviors:** In these studies, authors investigate in-game players' behaviors (e.g., persistence, engagement). By identifying these behaviors, we can group players according to different behaviors or simply check if a student shows a specific one. For example, Dicerbo [11] used evidence extracted from log files to create a measure of persistence. Similarly, Ventura and Shute [44] also created a measure of persistence, validating it against another existing measure and concluding that the GBA predicted students' learning.
- 3) **Assessment:** In these studies, games are used to report measures that aim to evaluate students. This allows for improvements in the learning process using this evaluation measure instead of classic evaluation methods or providing personalized feedback. In their work, Weiner and Sanchez [45] created an alternative measure using a virtual reality game that calculated scores to indicate specific cognitive abilities.
- 4) **Interventions:** Games can also be used to investigate the effect of some interventions while playing. For example, we can use feedback messages to notify the learner with positive (or negative) feedback to observe how this intervention influences its performance and behavior. Another typical example is switching the order of in-game elements or testing different game features. Cutumisu et al. [46] used a psychophysiological methodology to investigate attention allocation to different feedback valences (i.e., positive and negative feedback). With that purpose in mind, they used an eye tracker to collect accurate information about individuals' locus of attention when they process feedback.
- 5) **Framework proposal:** In these papers, the authors propose the design of a new framework to be used within the context of GBA. We can see an example in [47], where the authors examined the process of creating a Bayesian network framework through different techniques (e.g., using correlation matrixes, IRT) to create scoring models for assessing students.
- 6) **Game design proposal:** The authors provide a game design that can be used for assessment purposes. For example, Rivera and Suescún [48] show the design of an online GBA to help students improve their learning outcomes and promote the development of general and transferable skills, such as the ability to solve problems in complex situations, and working under pressure.

Some studies focused on more than one of the categories described above. For example, Weiner and Sanchez [45] used a virtual reality game to calculate a score measure for each student (assessment) and they proved that these calculated scores are the best used by comparing them to classic measures (GBA evaluation).

We can see the number of papers fitting each category in Fig. 6. GBA evaluation is the most common category (38 studies, 58.5%), followed by assessment (34 studies, 52.3%) and framework proposal (12 studies, 18.5%).

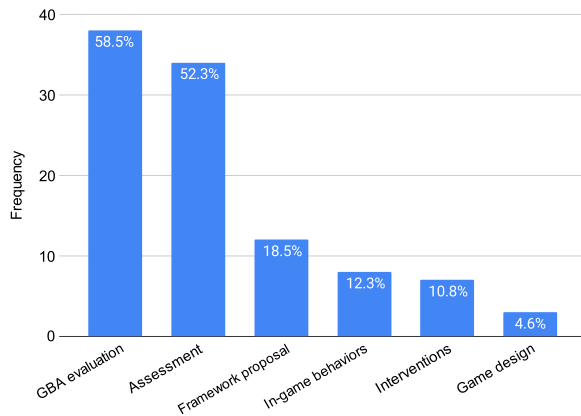


Fig. 6. Number of studies and rate in the collection per purpose of the research. More than one purpose is possible for a given paper.

The less common category is game design proposal, with only three papers fitting (4.6%). We can conclude that most papers focused on using games to assess learning, but they also tried to prove that this assessment was a valid measure to be used in real educational contexts.

C. What is the Domain of the GBA? (RQ3)

From reviewing the selected papers, we identify four major domain categories: STEM, humanities and social sciences, cognitive and soft skills, and physiological capacities. As some of the categories have more than one related area, we also consider some subcategories fitting into them. We describe each domain category as follows in detail.

- 1) **STEM:** In this category, we include papers that are related to science, technology, engineering, and mathematics. For example, Chiu and Hsieh [49] showed the different teaching methods of second-grade elementary students in fraction concepts (mathematics), while Kim et al. [26] aimed to assess the understanding of Newton's three laws of physics using a two-dimensional physics game.
- 2) **Humanities and social sciences:** Papers related to humanities and social science areas (e.g., art, music, language) fit in this category. As this is a wide area, we have also defined some subcategories to better categorize the papers. These subcategories are language, art, and history. Studies that do not fit into one of those three categories are categorized as other. As an example of the art category, we highlight the work in [50], where the authors used a game in which players collect data about the musical interests of an in-game character and use these data to make decisions about which artists to sign and what songs to record. We can see another example (related to language) in [51], where the researchers described the design of an argumentative reasoning task within a scenario-based assessment enhanced with game elements.

3) **Cognitive and soft skills:** Cognitive skills are the core skills your brain uses to think, read, learn, remember, reason, and pay attention [52]. Cognitive skills help to process new information by taking that information and distributing it into the appropriate areas in the brain. Developing cognitive skills helps to complete this process more quickly and efficiently, helping people to understand and effectively process new information [53]. Moreover, soft skills are described as a combination of interpersonal and social skills, including the ability to communicate, coordinate, work under pressure, and solve problems [54]. In this category, we consider attention, memory, logic and reasoning, visual processing and speed, and soft skills. We find papers that have measured interesting skills, such as [55], which included a series of reasoning activities to measure argumentation skills (which is related to logic and reasoning), or [56], [57], which used GBAs to assess candidates' soft skills.

4) **Physiological capacities:** Physiological functional capacity is the ability to perform the physical tasks of daily life and the ease with which these tasks can be performed. We could assess daily physical tasks, such as Rodríguez de Pablo et al. [58], who used a set of games to provide a fast, quantitative, and automatic evaluation of the arm movement function. Furthermore, other works focused on assessing mental abilities, such as motivating children with autism, to make more eye contact [59].

There are also papers fitting more than one category at once. For example, in [60] and [61], researchers used a GBA for measuring argumentation and pragmatic skills. This research measured language competencies (which is part of humanities and social sciences) but it also measured cognitive and soft skills. We can see the full tree showing the distribution of studies into categories and subcategories in Fig. 7.

The three predominant categories are cognitive and soft skills (28 studies, 43.1%), STEM (19 studies, 29.2%), and humanities and social sciences (17 studies, 26.2%). Taking a look at each subcategory, we note that the main area in STEM is science (10 studies, 15.4%). The main field in cognitive and soft skills is logic and reasoning, with 17 papers (26.2%), while in humanities and social sciences, the predominant subcategory is language, with nine papers (13.8%).

D. Is the Game/Tool Used Available to the Public? (RQ4)

A critical aspect of research is the availability of the results obtained to be used by the general public. It is essential to make tools accessible so that researchers can replicate experiments and practitioners can use them as part of their teaching. From our analysis, we find three primary categories: Currently available, Not available (NA), and Not specified.

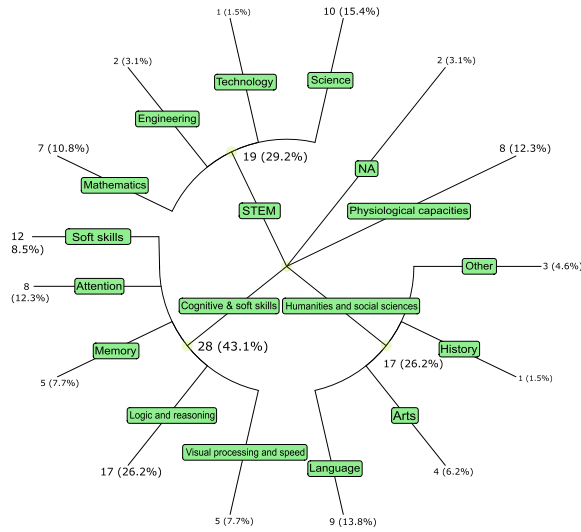


Fig. 7. Category tree for RQ3.

- 1) **Currently available:** The game/tool used in the corresponding research is currently available (using the web portal specified by the authors) for public use (e.g., [62], [63]).
- 2) **Not available:** The game/tool used in the research was presented as initially available in the paper, but currently, it is no longer accessible based on our attempt to access the site (e.g., [64], [65], [66]).
- 3) **Not specified:** Researchers did not specify the tool's availability; it is more than likely that it is not accessible (e.g., [67], [68], [69]).

Fig. 8 shows that most papers (51 studies, 78.5%) did not specify if the tool is accessible or not. Another minority of papers (9 studies, 13.8%) offered their tool publicly. The rest of the studies (five papers, 7.7%) initially offered their games but they are currently unavailable. In addition, we did not find any open-source game across the studies included in the review.

E. What is the Size of the Data Sample Used in the Study? (RQ5)

In this RQ, we classify the different data collections used in the studies based on their data sample size. Investigating the sample size is relevant since the use of larger data samples will allow better generalization of the research results, as well as the possibility of applying more complex algorithms (e.g., neural networks), which often require large amounts of data to outperform other models [17]. Although the sample size can be relevant for some aspects, such as preventing overfitting in some methods, it is not related to the study's rigor (i.e., using a larger sample does not make a study more rigorous). From the coding process, we present four categories.

- 1) **Fewer than 50 participants:** These papers involved fewer than 50 participants in their empirical studies. We find studies with small data samples, such as [70], using data from 30 postgraduate students, or [71],

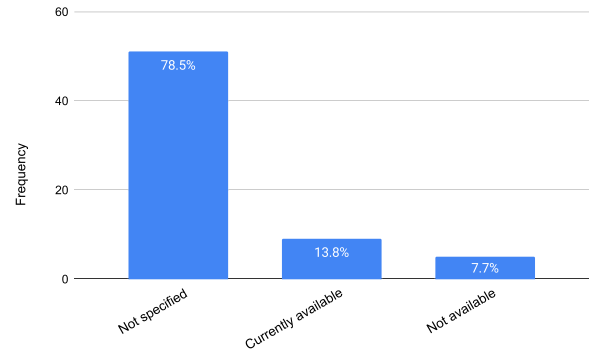


Fig. 8. Number of studies and rate in the collection per research availability.

which used a sample of 20 healthy controls patients and 18 patients with Alzheimer's disease to evaluate the usability of a tool created to assess cognitive functions.

- 2) **Between 50 and 250 participants:** These papers involved between 50 and 250 participants in their studies. For example, Leonardou et al. [72] used data from 77 primary school pupils for assessing and improving multiplication skills. We see another example in [73], which used data from 95 children from the final year in preschool to measure psychoacoustic thresholds.
- 3) **Between 250 and 500 participants:** These papers involved between 250 and 500 participants in their studies. For example, Gjicali et al. [74] used data from 433 students who played a game simulating an artificial culture with norms embodying two cultural concepts: hierarchy and collectivism.
- 4) **More than 500 participants:** These papers used data from more than 500 participants in their research. Hautala et al. [75] used data from 723 students to investigate reading difficulties, concluding that the GBA could be successfully used to identify students with reading difficulties with acceptable reliability (Cronbach's alpha 0.93 and 0.87). Some other studies used a huge sample, such as [12], which used data from 5545 students to measure engagement and cluster students to finally report four different engagement profiles.

Other papers (e.g., [76], [77]) did not specify the data sample size of the study and we categorize them as NA. Fig. 9 summarizes the results of the data sizes across papers. We can see that only 16 papers (24.6%) used more than 250 participants in their students, and only 9 papers (13.8%) used data from more than 500 students, meanwhile most of the papers (58.5%) used data from fewer than 250 participants. We also see that a significant amount of papers (11 studies, 16.9%) did not specify the data sample size in their studies.

F. What Computational Methods and Algorithms Have Been Applied in the Research? (RQ6)

After exploring the data samples that were retrieved across papers, our goal was to examine the methods that were applied for its analysis. We believe that the methods being applied are

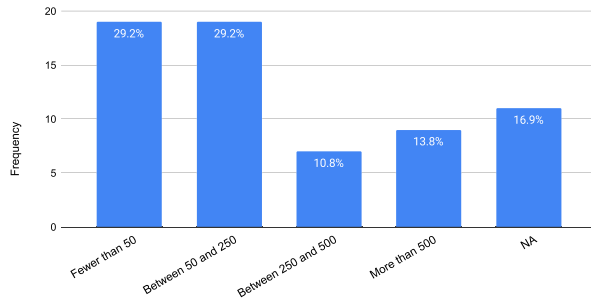


Fig. 9. Number of studies and rate in the collection per data sample size.

crucial since they are the link between the evidence generated by learners and the assessment. Accordingly, we identified five different groups of methods for analyzing the data: Descriptive statistics, Machine learning, Knowledge inference, Deep learning, and Sequence mining. Following is the description of each group in detail.

- 1) **Descriptive statistics:** These encompass further mathematical analyses covering various methods, tests, and visualizations. We identified several papers that applied summary statistics (e.g., mean, variances) [78], correlations [79], and visualizations [80].
- 2) **Machine learning:** It is a part of AI and covers a set of methods that allow systems to learn and improve from historical data automatically. We noted that the authors used two significant families of machine learning methods: supervised learning and unsupervised learning. Supervised learning includes techniques such as regression [81] while unsupervised learning uses other methods, such as clustering techniques like *k*-means [12] or dimensionality reduction techniques such as principal component analysis [82]. For example, Arce-Lopera and Perea [83] developed a game for evaluating the logic abilities of first-year university students. They tried to compare the measures obtained by paper-based tests with those obtained using the game by conducting a linear regression (which is a supervised method). The authors concluded that the measures obtained from both methods were not significantly different.
- 3) **Knowledge inference:** It refers to the acquisition of new knowledge from existing facts based on certain rules and constraints. One way of representing these rules and constraints is through the use of logic rules, formally known as knowledge representation [84]. Common knowledge inference methods that several studies have used are Bayesian networks [85] and fuzzy cognitive maps [67]. In [86], the researchers proposed a dynamic Bayesian network modeling approach for measuring student performance from an educational video game. The results supported the usefulness of Bayesian networks to characterize

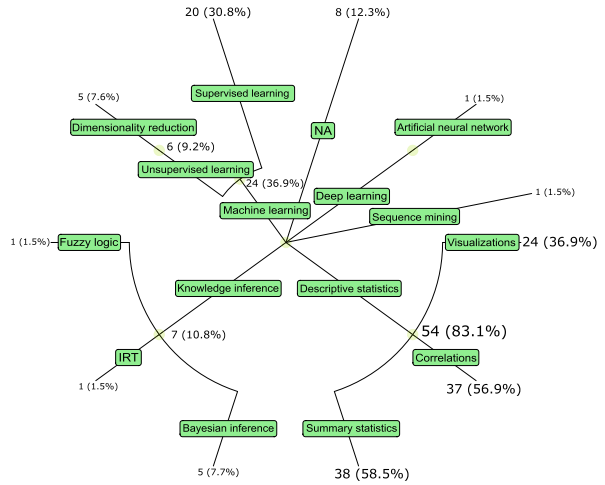


Fig. 10. Paper distribution based on the methods used.

and accumulate evidence regarding students in games and related assessment environments.

- 4) **Deep learning:** An artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for decision-making [87]. An example is the work of Chen et al. [88], who used long short-term memory (LSTM), an artificial recurrent neural network architecture.
- 5) **Sequence mining:** The objective of sequence mining is to unlock useful knowledge hidden in sequence data [89]. Specifically, Gomez et al. [90] used sequence mining to identify sequences and errors by transforming raw data into meaningful sequences that are interpretable and actionable for teachers.

In Fig. 10, we can see the different families of techniques and the number of papers that used them in their research. We can see that most papers (83.1%) used descriptive statistics, and almost none of them used deep learning (only one paper, 1.5%). We also noted that 83.3% of the papers that used machine learning techniques used supervised learning too, specifically, most of them used different types of regressions.

G. What Stakeholder is the Intended Recipient of the Research Results? (RQ7)

A stakeholder is defined as a person with an interest or concern in something, especially a business [91]. In our study, we consider the paper's stakeholder as the person to whom the results are directed, even though the paper's contribution might have other secondary stakeholders. Specifically, we have two main groups of stakeholders: researchers and the final user.

- 1) **Researchers:** If the paper's contribution is methodological, we expect that the paper's main stakeholders will be researchers. For example, Lonergan et al. [92] created a paper-based assessment and a GBA in order to measure students' performance, cognitive states, and

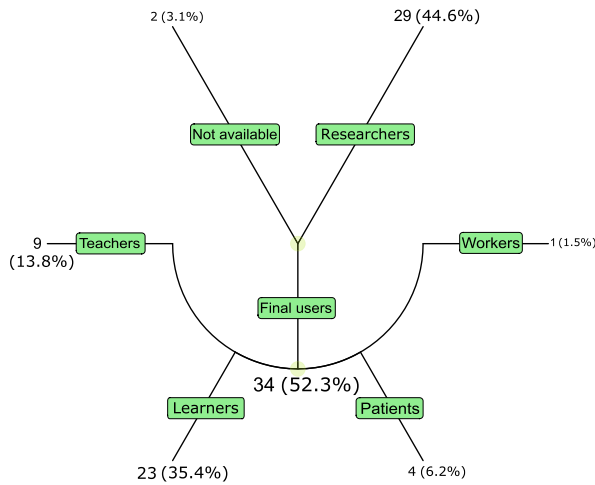


Fig. 11. Paper distribution tree based on the main stakeholder.

satisfaction related to both assessment methods. The authors concluded that smaller versatile GBAs may have a greater impact on the student's cognitive capabilities and could enhance student performances during, for example, a final exam or short formative assessments. Moreover, Tsai et al. [93] proposed an online learning system using different gaming modes of classic tic-tac-toe games to explore how different gaming modes and feedback types in this game-based formative assessment affect knowledge acquisition effectiveness and perceptions of participation.

- 2) **Final user:** If the paper's results are to be used by final users or are validating the GBA, we consider that the main stakeholder will be the final user in that context (e.g., teachers and students). In their work, Ciman et al. [94] designed a game to support children with cerebral visual impairment, developing a mobile version of the game to be used by children easily at home on any platform. Delcker and Ifenthaler [95] also developed a mobile app that makes an automated analysis of the data and provides information about children's language skills. Other papers focused on teachers, such as [96], where the authors used a GBA to develop a set of visualizations to support teachers in classrooms.

Fig. 11 shows the number of studies that focused their work on each of the stakeholders. We see that 34 studies (52.3%) were directed to final users, mainly students. Moreover, 29 studies (44.6%) focused on researchers as the main result recipients. Two further studies (e.g., [97]) did not provide results and we categorized them as NA. Focusing on studies directed to the final user, we see that the majority of papers are directed to learners (35.4%) and teachers (13.8%).

H. What Limitations and Challenges Do the Authors Address? (RQ8)

Limitations show potential weak points of the study that researchers usually highlight regarding their work such as

constraints in research design or methodology. We can group the limitations that the authors faced in the six following categories: game design, data sample, methodological, technical, integration, and validation.

- 1) **Game design:** An appropriate design of a game is crucial for learners' assessment since the GBA design must be adapted based on the constructs that will be evaluated. It requires a great effort to design a good GBA, aligning the evidence collected with the final purpose of the assessment. Designers might have different goals when developing a GBA [14]: "for game design, engagement; for instructional design, developing key concepts and capabilities in the target domain; for assessment design, evoking evidence of those capabilities for the intended use case(s)." Moreover, many game design decisions play a role in what kind of game performance is achieved and its meaning [98]. Future designers should consider these concerns to achieve better designs, thus creating more engagement and facilitating the development of the key concepts and capabilities intended for learning.
- 2) **Data sample:** Data were crucial for our review because GBA is based on the evidence, stored as data, generated by the students' interaction with the games. We examined each paper and found several limitations related to data. Jackson et al. [99] reported that they had a small sample size and that larger sample sizes would be necessary to detect smaller effects. We see a similar example in [100], where the authors had a sample collection of 67 students but only four of those 67 student samples were used in their empirical study. The work in [101] described the difficulty of designing a good data model as there are usually conflicts between programmers and assessment designers, usually complicated by constraints related to budgets and schedules.
- 3) **Methodological:** This category includes challenges and limitations related to the methods, algorithms, or techniques used. For example, Yu et al. [102] wanted to collect additional data to explore learners' behavioral patterns during gameplay. We see another example in [103] since the authors reported that the assessment developed in this study only includes a part of number sense (this term refers to a group of key math abilities), and, in order to complete the number-sense battery, the assessment tools for the other components of number sense are needed to be developed.
- 4) **Technical:** It is defined as a challenge involving how a machine or system works. This could include storage limitations, computing power, or even limitations related to sensors used in the study. Ibryamova and Stefanov [104] pointed out the necessity of a database (to store information about students' achievements) since they could not store that information as well as the necessity of an administrator module to facilitate developing and modifying game elements.
- 5) **Integration:** Incorporating game activities as part of the curriculum in schools remains limited due to

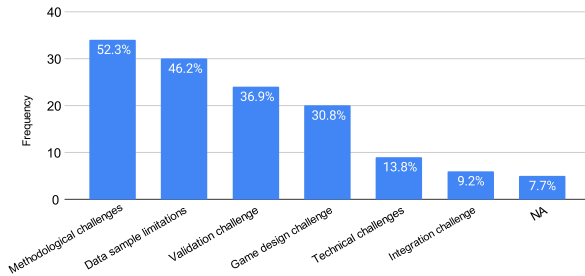


Fig. 12. Number of papers based on their limitations and challenges.

certain factors such as the schools' budget or the rigidity of subjects' classic curriculum. Halverson and Owen [64] claimed that if GBAs can show that games can serve as assessments that generate reliable evidence, we could then legitimize the potential of games and then break the social conventions that limit the potential of learning and assessment technologies in schools.

- 6) **Validation:** One of the most significant parts of the research is the validation of the results. Validation is intended to ensure that the proposed methods and the accomplished results proved satisfactory by conducting empirical experiments. Sanchez and Langer [105] suggested that the games used in their study were entertainment games, and further research could be oriented to validate their results with games designed for assessment purposes.

Some other studies did not report any challenges or limitations (e.g., [106]). Fig. 12 shows that methodological challenges are the most common ones (34 studies, 52.3%), followed by data sample limitations (30 studies, 46.2%). On the other hand, the least frequent limitations are related to integration (6 studies, 9.2%) and technical challenges (9 studies, 13.8%).

IV. DISCUSSION

In this section, we first present a summary and discussion of our main findings. We can also see a summary of these main findings in Table I. Next, we present a discussion of past and future challenges regarding games for assessment. Finally, we address existing limitations in this study and the implications of our research.

A. Current Trends

First of all, we analyzed the contexts where the studies were applied (RQ1), finding that most of them took place in K-16 education, especially in high school and middle school. This is an exciting finding because young kids and teenagers represent the major force whose 21st-century competencies development will be heavily impacted by technology [16], [107]. Moreover, children and adolescents are an ideal target since the familiarity of these users with gaming environments and game mechanics facilitates their interactions with games [17].

TABLE I
SUMMARY OF THE MAIN FINDINGS

| Research question | Categories | Count | % |
|------------------------------------|----------------------------------|-------|-------|
| Context (RQ1) | Formal education | 41 | 63.1% |
| | Medical | 8 | 12.3% |
| | Workplace | 5 | 7.7% |
| | Not Available | 11 | 16.9% |
| Main Purpose (RQ2) | GBA evaluation | 38 | 58.5% |
| | In-game behaviors | 8 | 12.3% |
| | Assessment | 34 | 52.3% |
| | Interventions | 7 | 10.8% |
| | Framework proposal | 12 | 18.5% |
| | Game design proposal | 3 | 4.6% |
| Domain (RQ3) | STEM | 19 | 29.2% |
| | Cognitive and soft skills | 28 | 43.1% |
| | Humanities and social sciences | 17 | 26.2% |
| | Physiological capacities | 8 | 12.3% |
| | Not Available | 2 | 3.1% |
| Availability (RQ4) | Not specified | 51 | 78.5% |
| | Currently available | 9 | 13.8% |
| | Not available | 5 | 7.7% |
| Sample Size (RQ5) | Fewer than 50 participants | 19 | 29.2% |
| | Between 50 and 250 participants | 19 | 29.2% |
| | Between 250 and 500 participants | 7 | 10.8% |
| | More than 500 participants | 9 | 13.8% |
| | Not available | 11 | 16.9% |
| Algorithms/techniques (RQ6) | Descriptive statistics | 54 | 83.1% |
| | Machine learning | 24 | 36.9% |
| | Deep learning | 1 | 1.5% |
| | Sequence mining | 1 | 1.5% |
| | Knowledge inference | 7 | 10.8% |
| | Not available | 8 | 12.3% |
| Stakeholder (RQ7) | Researchers | 29 | 44.6% |
| | Final user | 34 | 52.3% |
| | Not available | 2 | 3.1% |
| Limitations (RQ8) | Technical | 9 | 13.8% |
| | Game design | 20 | 30.8% |
| | Data sample | 30 | 46.2% |
| | Methodological | 34 | 52.3% |
| | Integration | 6 | 9.2% |
| | Validation | 24 | 36.9% |
| | Not available | 5 | 7.7% |

Regarding RQ2, we found that the majority of GBA studies focused on students' assessment and the validation of the game or the tool used. This suggests that having established that games are helpful for other purposes beyond entertainment, there is an increasing interest in using games as a natural alternative to classic evaluation methods, validating and comparing them against those traditional alternatives. Moreover, the fact that researchers also focused on the validation of the GBA used is a promising finding. Specifically, Gris and Bengtson [18] pointed out the lack of evidence about engagement and usability needs, especially with well-assessed reliability and validity. We also noticed that few studies had the main purpose of proposing or validating a game design for assessment. Although many studies proved that GBA could improve students' learning outcomes, we should not forget game design. The literature reveals that game design is essential, and several distinctive design elements, such as narrative context, rules, goals, rewards, multisensory cues, and interactivity, seem necessary to stimulate the desired outcomes [108], [109].

We also extracted four predominant domains (RQ3) across studies. A large proportion of the analyzed studies aimed at practicing and assessing content related to STEM as well as humanities and social sciences. This is not surprising since many of the studies took place in schools and high schools, and

the use of games in these contexts is an ideal opportunity to teach content related to the main subjects at those ages. Another large number of papers also focused on developing and measuring cognitive and soft skills. Using game design as a context to teach higher order thinking skills has drawn attention from researchers since schools usually place heavy emphasis on covering and delivering content knowledge [110]. Moreover, this could be useful not only in educational contexts, as we have seen some studies that measure cognitive skills for medical purposes (e.g., rehabilitation) [68], [71]. However, the researchers in [111] pointed out the lack of research on 21st-century skills such as creativity and critical thinking.

We discovered that many of the studies had small data samples (RQ5). Furthermore, a significant part of the studies did not specify the data sample size used in the experiment. This is also noticed by the researchers themselves, as nearly half of the studies reported data sample limitations. Moreover, apart from collecting the sample size, we also tried to collect information about the type of data collected. However, almost no study included information related to the type or format of the data used.

Across papers, researchers used many different algorithms and techniques (RQ6) to analyze the data. We classified them into five categories and found that the most common ones are descriptive statistics and machine learning. With that said, we note that the majority of papers used statistical analyses or basic machine learning algorithms, and few studies used more complex or advanced methods, which might be more adequate to model students' knowledge properly. However, those techniques that are easier to implement are also the ones chosen more frequently by researchers. Therefore, more work is needed to develop specialized GBA methods that are also affordable to implement. This could perhaps be done through open-source libraries and more reproducible research. Moreover, we consider that making results interpretable is an essential part of the assessment, and one way to reach this interpretability is by using visualizations. Visualizations are essential components of research presentation and communication because of their ability to represent large amounts of data [112] and because it is easier for the brain to comprehend an image as opposed to words or numbers [113]. We think this is a promising way to integrate games in schools, and we realized that studies now tend to use visualizations to communicate their results (e.g., [62], [80]).

Finally, we wish to report the scarce information regarding games and tools availability (RQ4). The majority of studies did not provide any information on how to access or use their tools. In addition, some studies made tools public but expired, being inaccessible at present. This underscores the low transference of this research to practice, and thus, we encourage authors to make their products and results publicly accessible since we consider that this is an essential part of this type of research.

B. Open Challenges

From our results and previous related reviews, we find some open challenges in the area that authors usually report. A description of each of these challenges is found as follows.

Chin, Dukes, and Gamson [114] address the challenge of how to make appropriate assessments. They noted that pretest and post-test measures are a good manner to make an assessment, and they also recommended unobtrusive ways to collect data, such as another person taking notes during game-play. In our review, we noted that, at present, most studies found an appropriate method to make good assessments using evidence-centered design (ECD) and stealth assessment. ECD framework views assessment as an evidentiary argument, that is, an argument from which we observe what students say, do, or make in a few particular circumstances [115]. Moreover, stealth assessment represents a unobtrusive, yet powerful process by which learner performance data are continuously gathered during playing and learning, and inferences are made about the level of relevant competencies, maintaining the learners' flow and engagement [116]. Since ECD and stealth assessment are two common practices in current research, we could claim that the objective of making appropriate assessments using unobtrusive methods has been accomplished.

What data are going to be collected is as important as how to collect these data, and another present challenge is the design of games for specific assessment purposes. Akcaoglu and Koehler [110] indicated that games that present a hidden questionnaire format do not engage learners, while well-designed games can engage learners in reflective thinking [117]. Although we identified a few papers with the main objective of providing a good game design, many of them have developed an excellent game for other purposes. Some examples are given in [12], [71], and [73]. Future research should focus on complex game designs rather than the typical simple quiz design, employing multiple game-design elements such as collaboration, role-playing, narrative, exploration, and complexity [16].

An important open challenge at present is replication and transferring the research to practice. In addition to the findings in our literature review about the game or tool being unavailable in most cases, All et al. [118] mentioned replication issues with certain studies due to missing information in multiple areas of the study. It is crucial to provide a detailed description of the procedure followed to conduct the study. While the community is currently demanding more standardized open science practices, this problem is still present currently. Besides, Alonso-Fernandez et al. [17] noted that most papers did not describe the format in which they collected the data, so we cannot know if they used a standard or relied on their data formats, which represent even more replication and reusability issues. In addition, having open-source games or tools would be especially helpful for researchers. Unfortunately, we did not find any available open-source games across the studies. This problem of missing information is a familiar issue in multiple research fields (nearly every field is affected), and it leads to other problems such as low reproducibility. In fact, the terms "reproducibility crisis" and "replication crisis" have gained significant popularity over the last decade [119]. To fix this issue, the community is demanding more preregistered studies, open data, open analyses, and open access publications [120], and this can be systematized by the guidelines of the publishers, governments, and research communities [121].

Regarding the methods and techniques, we identify the challenge of implementing learner modeling algorithms. As we mentioned above, researchers usually use simple techniques to conduct their studies. In addition to Alonso-Fernandez et al. [17] noting that limitation, we confirmed it in our results. In our review, 52.3% of the papers reported methodological challenges to be addressed in future research, most of them related to the use of more complex metrics and techniques to infer new information. It is important to benefit from more advanced techniques (e.g., knowledge inference techniques, deep learning techniques) that can allow us to infer more complex and valuable information from the data collected. However, an important limitation of many of those advanced techniques is their low interpretability. Even if visualizations are a promising way to improve the presentation of results and communication, they cannot improve the model's interpretability themselves. According to the researchers in [122], with machine learning models being increasingly used, there has been an interest in developing interpretable models. However, there have been relatively few experimental studies investigating whether these models achieve their intended effects. Thus, the development of new models to provide better interpretability in GBA environments and their validation is still an open challenge.

We found several studies that described data sample and validation challenges. Since most evaluations are conducted with small samples, typically corresponding to one classroom's size, these studies present low statistical power, having a reduced chance of detecting actual effects [123]. Thus, studies must use larger data samples to improve the results' generalization and validity. However, collecting large samples of in-context data is also a cumbersome task. Finally, a few empirical studies discussed the challenge of implementing GBA in the classroom, but this is a significant problem. Many teachers are still unsure about how to integrate game activities with the regular curriculum, and it is crucial to provide guidelines that can facilitate teachers to deploy games in the classroom more easily and flexibly [124].

C. Limitations and Implications

This review is mainly limited by the paper selection. First of all, we have only used the key term "game-based assessment" to perform our document search, based on the papers' keywords and titles. However, other communities could also be working on games for assessment purposes, but they might be using slightly different terms to describe their work. Therefore, those studies might not be included in our review. Nevertheless, we purposely opted for this term to analyze the core of GBA while also having a manageable selection of papers for this study. Furthermore, we focused our attention on Scopus and the Web of Science, the two primary academic databases. However, there could be other peer-reviewed academic papers indexed in different databases, as well as non-peer-reviewed publications including preprints, technical or white reports that could be missing in our review, and also nonacademic work being conducted in industrial companies and by practitioners.

Regarding the computational methods and algorithms used, we have identified a set of categories based on the qualitative review of each selected paper. However, there might be studies using less quantitative approaches that might be missing in this review due to the review methodology itself. Finally, we have based our RQ generation on a simplified process that involves the general steps required in GBA projects, but there might be other potential and valuable RQs about the GBA field missing in this review.

We found that most studies emphasized GBA implementation and comparisons between games and classic assessment methods. More studies are needed to systematically develop and improve game design, adopting design-based research methods, as mentioned in [125]. The potential of GBA is now emerging, coinciding with the rise of Big Data. Data mining and visualization techniques on player interaction logs can provide different stakeholders with valuable insights into how players interact with the game [126]. The increasing interest in games as a learning tool also indicates their potential as actual assessment tools. In our review, we found that GBAs are not only being used in K-16 education but also in medical and professional areas, among others. As expected, the most frequent area where GBAs are being applied is K-16 education since children and adolescents are the leading groups whose development will be affected by technology.

Despite this dominating use in education, we can see the great potential that GBAs have in many other contexts. Concerning the professional environment, companies have begun to include assessment games for the recruitment of staff and the selection process. This is a relatively new trend due to certain limitations, such as the cross-domain generalizability of behaviors between game and workforce contexts, which needs further research [40]. In medical environments, the use of GBAs can also be helpful for multiple purposes. Some examples are the possibility of recreating a virtual environment with daily life activities, allowing a precise and complete cognitive evaluation, which can be useful to treat certain diseases such as Alzheimer's [71] or using games to rehabilitate children with cerebral visual impairment using an eye-tracker [94]. Due to the above, we firmly believe that the future of games for assessment is promising; however, further research is needed to overcome the existing problems, and increase the still limited application of games in real-life environments, in order to start building the classrooms of the future.

V. CONCLUSION

Technology is changing and improving every day, and this is also making a significant impact on educational areas. Moreover, playing games is one of the most popular activities in the world, and the technological revolution that we are experiencing allows the implementation of games as alternative assessment tools in educational environments. However, previous studies suggest that the use of games also presents some challenges, such as finding the time for both the presenter/instructor and student to learn the systems employed, the financial impact on both parties, and technical limitations [1], [6]. We can tackle

all these challenges by facing current limitations and revealing the great potential games have for assessment. In this article, we represent a novel analysis and the first literature review of the emerging research field of GBA. Its main purpose was to review empirical studies of digital GBAs published until 2020. Based on a detailed systematic review of the 65 selected papers, we concluded that games are mainly used in K-16 education for assessment and validation purposes. The domain of the games used is usually related to STEM and cognitive skills but other domains emerged from our analysis, such as social sciences and physiological capacities. Moreover, we note that, although few GBA studies had the purpose of proposing an adequate game design for assessment, most studies used games designed specifically for assessment purposes, employing complex game-design elements such as collaboration, narrative, or role-playing. In addition, we found that most of the studies used small data samples and simple techniques to process these data and assess students. Finally, we found that most of the studies do not provide public access to their tools, or they overlook links and let them expire over time, which makes it impossible to reproduce the results or even try their game.

Future work should address the current challenges emerging from our review, as those are the main barriers to actual systematic adoption of games for assessment. For example, the next generation of GBA studies should ensure that enough data are collected to have meaningful and reliable results since one of the main limitations of the current research was the size of the data sample collected. Moreover, they should also address the game design that will be used for assessment, as many studies use games designed for other purposes (e.g., entertainment) and overlook the vital link between the design of a game and collecting the necessary evidence for the assessment. In that sense, it would be good to work on conceptual GBA pieces or frameworks that can provide a set of guidelines for the design. Moreover, classic performance indicators such as completion times or scores could still be included in future studies, but GBA also needs to apply more specific and complex algorithms (e.g., knowledge inference or deep learning techniques) specifically designed for learner modeling and assessment purposes. The use of more complex techniques, along with larger data samples, could substantially improve the reliability and generalization of the results. We also believe that future studies should continue exploring the use of visualizations and dashboards to integrate games in schools, adopting a more intuitive approach rather than providing teachers with raw numerical outputs or metrics, which are usually harder to understand. Teachers should also have a more important role in future work to address digital and assessment literacy issues, as well as the potential interpretability and actionability of GBAs. Finally, there are no theoretical frameworks within the GBA area (a related one regarding serious games could be [127]). Considering this lack of theoretical papers focused on describing GBA foundations, we believe that future work should address publications with additional content on the theoretical side.

Therefore, further research is needed to overcome current limitations and to continue exploring the possibilities of games

as assessment tools in other contexts and environments. Finally, we encourage authors to document their research in a reproducible and verifiable way, using beneficial open science practices by preregistering their study, sharing data and code for replication purposes, and if possible open sourcing the GBA tools with clear descriptions so that they can be used by interested stakeholders and researchers.

REFERENCES

- [1] L. S. Eiland and T. J. Todd, "Considerations when incorporating technology into classroom and experiential teaching," *J. Pediatr. Pharmacol. Therapeutics*, vol. 24, no. 4, pp. 270–275, 2019.
- [2] S. De Freitas, "Are games effective learning tools? A review of educational games," *J. Educ. Technol. Soc.*, vol. 21, no. 2, pp. 74–84, 2018.
- [3] ESA, "2020 essential facts about the computer and video game industry," *Entertainment Softw. Assoc.*, Tech. Rep., 2020.
- [4] ISFE, "ISFE key facts 2020," ISFE, Tech. Rep., 2020.
- [5] S. Papadakis, "The use of computer games in classroom environment," *Int. J. Teach. Case Stud.*, vol. 9, no. 1, pp. 1–25, 2018.
- [6] S. de Klerk and P. M. Kato, "The future value of serious games for assessment: Where do we go now?," *J. Appl. Testing Technol.*, vol. 18, no. S1, pp. 32–37, 2017.
- [7] D. Clow, "An overview of learning analytics," *Teach. Higher Educ.*, vol. 18, no. 6, pp. 683–695, 2013.
- [8] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, 2013.
- [9] Y. J. Kim and V. J. Shute, "Opportunities and challenges in assessing and supporting creativity in video games," in *Video Games and Creativity*. Amsterdam, The Netherlands: Elsevier, 2015, pp. 99–117.
- [10] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth, "Digital games, design, and learning: A systematic review and meta-analysis," *Rev. Educ. Res.*, vol. 86, no. 1, pp. 79–122, 2016.
- [11] K. Dicerbo, "Game-based assessment of persistence," *Educ. Technol. Soc.*, vol. 17, no. 1, pp. 17–28, 2013.
- [12] J. Ruiperez-Valiente, M. Gaydos, L. Rosenheck, Y. Kim, and E. Klopfer, "Patterns of engagement in an educational massively multiplayer online game: A multidimensional view," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 648–661, Oct./Dec. 2020.
- [13] D. Ifenthaler, D. Eseryel, and X. Ge, "Assessment for game-based learning," in *Assessment in Game-Based Learning*. Berlin, Germany: Springer, 2012, pp. 1–8.
- [14] R. Mislevy et al., "Psychometrics and game-based assessment," in *Technology and Testing*. Oxfordshire, U.K.: Routledge, 2015.
- [15] Y. J. Kim and D. Ifenthaler, "Game-based assessment: The past ten years and moving forward," in *Game-Based Assessment Revisited*. Berlin, Germany: Springer, 2019, pp. 3–11.
- [16] M. Qian and K. R. Clark, "Game-based learning and 21st century skills: A review of recent research," *Comput. Hum. Behav.*, vol. 63, pp. 50–58, 2016.
- [17] C. Alonso-Fernandez, A. Calvo-Morata, M. Freire, I. Martinez-Ortiz, and B. Fernández-Manjón, "Applications of data science to game learning analytics data: A systematic literature review," *Comput. Educ.*, vol. 141, 2019, Art. no. 103612.
- [18] G. Gris and C. Bengtson, "Assessment measures in game-based learning research: A systematic review," *Int. J. Serious Games*, vol. 8, no. 1, pp. 3–26, 2021.
- [19] X. Guan, C. Sun, G.-j. Hwang, K. Xue, and Z. Wang, "Applying game-based learning in primary education: A systematic review of journal publications from," in *Interactive Learning Environments*, vol. 30. Oxfordshire, U.K.: Taylor & Francis, 2022, pp. 1–23.
- [20] P.-Y. Chen, G.-J. Hwang, S.-Y. Yeh, Y.-T. Chen, T.-W. Chen, and C.-H. Chien, "Three decades of game-based learning in science and mathematics education: An integrated bibliometric analysis and systematic review," *J. Comput. Educ.*, vol. 9, pp. 455–476, 2021.
- [21] M. J. Page et al., "PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021, Art. no. n160. [Online]. Available: <https://www.bmj.com/content/372/bmj.n160>
- [22] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021, Art. no. n71.
- [23] K. S. Tekinbas and E. Zimmerman, *Rules of Play: Game Design Fundamentals*. Cambridge, MA, USA: MIT Press, 2003.

- [24] B. Kang et al., "Interactive games: Intrinsic and extrinsic motivation, achievement, and satisfaction," *J. Manage. Strategy*, vol. 5, no. 4, pp. 110–116, 2014.
- [25] S. De Freitas, *Learning in Immersive Worlds: A Review of Game-Based Learning*. Bristol, U.K.: Jisc, 2006.
- [26] Y. Kim, R. Almond, and V. Shute, "Applying evidence-centered design for the development of game-based assessments in physics playground," *Int. J. Testing*, vol. 16, no. 2, pp. 142–163, 2016.
- [27] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining" gamification," in *Proc. 15th Int. Acad. MindTrek Conf., Envisioning Future Media Environments*, 2011, pp. 9–15.
- [28] H. Al Fatta, Z. Maksom, and M. H. Zakaria, "Game-based learning and gamification: Searching for definitions," *Int. J. Simul., Syst., Sci. Technol.*, vol. 19, no. 6, 2018, Art. no. 41.
- [29] I. Ghergulescu and C. H. Muntean, "Measurement and analysis of learner's motivation in game-based e-learning," in *Assessment in Game-Based Learning*. Berlin, Germany: Springer, 2012, pp. 355–378.
- [30] J. A. Ruipérez-Valiente, "Unveiling the potential of learning analytics in game-based learning: Case studies with a geometry game," in *Handbook of Research on Promoting Economic and Social Development Through Serious Games*. Hershey, PA, USA: IGI Global, 2022, pp. 524–544.
- [31] A. Aghaei Chadegani et al., "A comparison between two main academic literature collections: Web of science and Scopus databases," *Asian social Sci.*, vol. 9, no. 5, pp. 18–26, 2013.
- [32] Elsevier, "About Scopus," 2021. [Online]. Available: <https://www.elsevier.com/es-es/solutions/scopus>
- [33] Clarivate, "Web of Science," 2021. [Online]. Available: <https://clarivate.com/webofsciencelibrary/solutions/web-of-science/>
- [34] A. Medelyan, "Coding qualitative data: How to code qualitative research," 2021. [Online]. Available: <https://getthematic.com/insights/coding-qualitative-data/>
- [35] M. J. Gomez, J. Ruipérez-Valiente, and F. J. García Clemente, "Supplementary materials: A systematic literature review of game-based assessment empirical studies: Current trends and open challenges," Accessed: Nov. 28, 2021. [Online]. Available: https://osf.io/34jk9/?view_only=865f94046fa84c45a013d986bb1c4f87.
- [36] K. Di Cerbo et al., "An application of exploratory data analysis in the development of game-based assessments," in *Serious Games Analytics*. Berlin, Germany: Springer, 2015.
- [37] J.-P. Lindenmayer et al., "Assessing instrumental activities of daily living (iADL) with a game-based assessment for individuals with schizophrenia," *Schizophrenia Res.*, vol. 223, pp. 166–172, 2020.
- [38] S. Wiloth, N. Lemke, C. Werner, and K. Hauer, "Validation of a computerized, game-based assessment strategy to measure training effects on motor-cognitive functions in people with dementia," *JMIR Serious Games*, vol. 4, no. 2, 2016, Art. no. e5696.
- [39] A. B. Collmus, M. B. Armstrong, and R. N. Landers, "Game-thinking within social media to recruit and select job candidates," in *Social Media in Employee Selection and Recruitment*. Berlin, Germany: Springer, 2016, pp. 103–124.
- [40] E. Short and N. Weidner, "Gamers at work: Predicting workplace-relevant behaviours across domains," *J. Gaming Virtual Worlds*, vol. 11, no. 2, pp. 161–177, 2019.
- [41] C. G. I. A. Stanciu and A. T. D. F. Stănescu, "Development of an integrated game based assessment approach—The next generation of psychometric testing," *Eur. J. Sustain. Develop.*, vol. 8, no. 5, pp. 270–270, 2019.
- [42] H. Hummel, D. Joosten-ten Brinke, R. Nadolski, and L. Baartman, "Content validity of game-based assessment: Case study of a serious game for ICT managers in training," *Technol., Pedagogy Educ.*, vol. 26, no. 2, pp. 225–240, 2017.
- [43] A. Marengo and A. Pagano, "Innovative ways to assess soft-skills: The in-basket game online experience," in *Proc. Eur. Conf. e-Learn. Acad. Conf. Int. Limited*, 2020, pp. 325–334.
- [44] M. Ventura and V. Shute, "The validity of a game-based assessment of persistence," *Comput. Hum. Behav.*, vol. 29, no. 6, pp. 2568–2572, 2013.
- [45] E. Weiner and D. Sanchez, "Cognitive ability in virtual reality: Validity evidence for VR game-based assessments," *Int. J. Selection Assessment*, vol. 28, no. 3, pp. 215–235, 2020.
- [46] M. Cutumisu et al., "Eye tracking the feedback assigned to undergraduate students in a digital assessment game," *Front. Psychol.*, vol. 10, 2019, Art. no. 1931.
- [47] R. Almond, "Tips and tricks for building Bayesian networks for scoring game-based assessments," in *Proc. Workshop Adv. Methodologies Bayesian Netw.*, 2015, vol. 9505, pp. 250–263.
- [48] L. Rivera and C. Suescún, "Game-based assessment for radiofrequency circuits courses in engineering," in *Proc. Front. Educ. Conf.*, 2015, pp. 1–5.
- [49] F.-Y. Chiu and M.-L. Hsieh, "Role-playing game based assessment to fractional concept in second grade mathematics," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 13, no. 4, pp. 1075–1083, 2017.
- [50] S. Basu, B. Disalvo, D. Rutstein, Y. Xu, J. Roschelle, and N. Holbert, "The role of evidence centered design and participatory design in a playful assessment for computational thinking about data," in *Proc. Annu. Conf. Innov. Technol. Comput. Sci. Educ.*, 2020, pp. 985–991.
- [51] Y. Song and J. Sparks, "Measuring argumentation skills through a game-enhanced scenario-based assessment," *J. Educ. Comput. Res.*, vol. 56, no. 8, pp. 1324–1344, 2019.
- [52] Learningrx, "What are cognitive skills?" 2021. [Online]. Available: <https://www.learningrx.com/what-is-brain-training-/what-are-cognitive-skills/>
- [53] I. E. Team, "Cognitive skills: What they are and how to improve them," 2021. [Online]. Available: <https://www.indeed.com/career-advice/career-development/cognitive-skills-how-to-improve-them>
- [54] J. Dixon, C. Belnap, C. Albrecht, and K. Lee, "The importance of soft skills," *Corporate Finance Rev.*, vol. 14, no. 6, pp. 35–38, 2010.
- [55] Y. Song and J. Sparks, "Building a game-enhanced formative assessment to gather evidence about middle school students' argumentation skills," *Educ. Technol. Res. Develop.*, vol. 67, no. 5, pp. 1175–1196, 2019.
- [56] I. Nikolaou, K. Georgiou, and V. Kotsasarlidou, "Exploring the relationship of a gamified assessment with performance," *Spanish J. Psychol.*, vol. 22, pp. 1–10, 2019.
- [57] E. M. Mosalam, G. A. El Khayat, S. Lazem, L. Cheniti-Belcadhi, and B. Said, "Assessing modelling readiness in a games environment," in *Proc. 7th Int. Conf. ICT Accessibility*, 2019, pp. 1–6.
- [58] C. Rodríguez-de Pablo, A. Savić, and T. Keller, "Game-based assessment in upper-limb post-stroke telerehabilitation," *Biosyst. Biobot.*, vol. 15, pp. 413–417, 2017.
- [59] V. Korhonen, H. Rätty, and E. Kärnä, "A pilot study: A computer game-based assessment of visual perspective taking of four children with autism with high support needs," *Scand. J. Disabil. Res.*, vol. 19, no. 4, pp. 281–294, 2017.
- [60] G. Tanner Jackson, B. Lehman, and L. Grace, "Awkward Annie: Impacts of playing on the edge of social norms," in *Proc. ACM Int. Conf. Found. Digit. Games Ser.*, 2020, Paper 73.
- [61] G. Jackson, L. Grace, P. Inglese, J. Wain, and R. Hone, "Awkward Annie: Game-based assessment of english pragmatic skills," in *Proc. Int. Conf. Adv. Comput. Entertainment*, 2018, vol. 10714, pp. 795–808.
- [62] H. Song, D.-J. Yi, and H.-J. Park, "Validation of a mobile game-based assessment of cognitive control among children and adolescents," *PLoS One*, vol. 15, no. 3, 2020, Art. no. e0230498.
- [63] H. Ketamo and K. Devlin, "Replacing pisa with global game based assessment," in *Proc. Eur. Conf. Games-Based Learn.*, 2014, pp. 258–264.
- [64] R. Halverson and V. Owen, "Game-based assessment: An integrated model for capturing evidence of learning in play," *Int. J. Learn. Technol.*, vol. 9, no. 2, pp. 111–138, 2014.
- [65] O. Gaggi, T. Sgaramella, L. Nota, M. Bortoluzzi, and S. Santilli, "A serious games system for the analysis and the development of visual skills in children with CVI: A pilot study with kindergarten children," in *Proc. Int. Conf. Smart Objects Technol. Social Good*, 2017, vol. 195, pp. 155–165.
- [66] M. Bertling, G. Tanner Jackson, A. Oranje, and V. Owen, "Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2015, vol. 9112, pp. 545–549.
- [67] H. Barón, R. Crespo, J. Pascual Espada, and O. Martínez, "Assessment of learning in environments interactive through fuzzy cognitive maps," *Soft Comput.*, vol. 19, no. 4, pp. 1037–1050, 2015.
- [68] M. Loachamin-Valencia, M.-C. Juan, M. Mendez-Lopez, and E. Perez-Hernandez, "Auditory and spatial assessment in inattentive children using smart devices and gesture interaction," in *Proc. IEEE 17th Int. Conf. Adv. Learn. Technol.*, 2017, pp. 106–110.
- [69] S. Pouzevara, S. Powers, G. Moore, C. Strigel, and K. McKnight, "Assessing soft skills in youth through digital games," in *Proc. Int. Conf. Eng. Res. Innovations*, 2019, pp. 3057–3066.

- [70] A. Mavridis and T. Tsiatsos, "Game-based assessment: Investigating the impact on test anxiety and exam performance," *J. Comput. Assist. Learn.*, vol. 33, no. 2, pp. 137–150, 2017.
- [71] V. Vallejo et al., "Evaluation of a novel serious game based assessment tool for patients with Alzheimer's disease," *PLoS One*, vol. 12, no. 5, 2017, Art. no. e0175999.
- [72] A. Leonardou, M. Rigou, and J. Garofalakis, "Techniques to motivate learner improvement in game-based assessment," *Information*, vol. 11, no. 4, 2020, Art. no. 176.
- [73] V. Abeele, J. Wouters, P. Ghesquière, A. Goeleven, and L. Geurts, "Game-based assessment of psychoacoustic thresholds: Not all games are equal!," in *Proc. Annu. Symp. Comput. Hum. Interact. Play*, 2015, pp. 331–342.
- [74] K. Gjicali, B. Finn, and D. Hebert, "Effects of belief generation on social exploration, culturally-appropriate actions, and cross-cultural concept learning in a game-based social simulation," *Comput. Educ.*, vol. 156, 2020, Art. no. 103959.
- [75] J. Hautala, R. Heikkilä, L. Nieminen, V. Rantanen, J.-M. Latvala, and U. Richardson, "Identification of reading difficulties by a digital game-based assessment technology," *J. Educ. Comput. Res.*, vol. 58, no. 5, pp. 1003–1028, 2020.
- [76] Y. Kim and V. Shute, "The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment," *Comput. Educ.*, vol. 87, pp. 340–356, 2015.
- [77] J. Perry, S. Balasubramanian, C. Rodríguez-De-Pablo, and T. Keller, "Improving the match between ability and challenge: Toward a framework for automatic level adaptation in game-based assessment and training," in *Proc. IEEE 13th Int. Conf. Rehabil. Robot.*, 2013, pp. 1–6.
- [78] M. Gómez-álvarez, J. Echeverri, and L. González-Palacio, "Games-based assessment strategy: Case systems engineer of Universidad de Medellín [estrategia de evaluación basada en juegos: Caso ingeniería de sistemas Universidad de Medellín]," *Ingeniare*, vol. 25, no. 4, pp. 633–642, 2017.
- [79] Y. Jaffal and D. Wloka, "Employing game analytics techniques in the psychometric measurement of game-based assessments with dynamic content," *J. e-Learn. Knowl. Soc.*, vol. 11, no. 3, pp. 101–115, 2015.
- [80] M. Cutumisu, D. Chin, and D. Schwartz, "A digital game-based assessment of middle-school and college students' choices to seek critical feedback and to revise," *Brit. J. Educ. Technol.*, vol. 50, no. 6, pp. 2977–3003, 2019.
- [81] D. Chin, K. Blair, and D. Schwartz, "Got game? A choice-based learning assessment of data literacy and visualization skills," *Technol., Knowl. Learn.*, vol. 21, no. 2, pp. 195–210, 2016.
- [82] C. Forsyth, T. Jackson, D. Hebert, B. Lehman, P. Inglese, and L. Grace, "Striking a balance: User-experience and performance in computerized game-based assessment," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2017, pp. 502–505.
- [83] C. Arce-Lopera and A. Perea, "Logic evaluation through game-based assessment," *Adv. Intell. Syst. Comput.*, vol. 973, pp. 243–250, 2020.
- [84] L. Tari, *Knowledge Inference*. New York, NY, USA: Springer, 2013, pp. 1074–1078.
- [85] V. Shute and L. Wang, "Assessing and supporting hard-to-measure constructs in video games," in *The Handbook of Cognition and Assessment*. New York, NY, USA: Wiley, 2016.
- [86] R. Levy, "Dynamic Bayesian network modeling of game-based diagnostic assessments," *Multivariate Behav. Res.*, vol. 54, no. 6, pp. 771–794, 2019.
- [87] M. Hargrave, "Deep learning," 2020. [Online]. Available: <https://www.investopedia.com/terms/d/deep-learning.asp/>
- [88] F. Chen, Y. Cui, and M.-W. Chu, "Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study," *Int. J. Artif. Intell. Educ.*, vol. 30, no. 3, pp. 481–503, 2020.
- [89] G. Dong and J. Pei, *Sequence Data Mining*, vol. 33. Berlin, Germany: Springer, 2007.
- [90] M. J. Gomez, J. A. Ruipérez-Valiente, P. A. Martinez, and Y. J. Kim, "Exploring the affordances of sequence mining in educational games," in *Proc. 8th Int. Conf. Technological Ecosyst. Enhancing Multiculturality*, 2020, pp. 648–654.
- [91] Oxford, "Oxford languages and Google," 2021. [Online]. Available: <https://languages.oup.com/google-dictionary-en/>
- [92] M. Loneragan, L. De Wet, and A. Burger, "Technology as a tool to improve understanding of assessment questions," *Adv. Intell. Syst. Comput.*, vol. 1217, pp. 250–256, 2020.
- [93] F.-H. Tsai, C.-C. Tsai, and K.-Y. Lin, "The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment," *Comput. Educ.*, vol. 81, pp. 259–269, 2015.
- [94] M. Ciman, O. Gaggi, T. Sgararella, L. Nota, M. Bortoluzzi, and L. Pinello, "Serious games to support cognitive development in children with cerebral visual impairment," *Mobile Netw. Appl.*, vol. 23, no. 6, pp. 1703–1714, 2018.
- [95] J. Delcker and D. Ifenthaler, "Mobile game-based language assessment," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 3, pp. 195–206, 2020.
- [96] P. Martínez, M. Gómez, J. Ruipérez-Valiente, G. Pérez, and Y. Kim, "Visualizing educational game data: A case study of visualizations to support teachers," in *Proc. CEUR Workshop*, 2020, vol. 2671, pp. 97–111.
- [97] J. Paiva and J. Leal, "Asura: A game-based assessment environment for Mooshak," in *Proc. 7th Symp. Lang., Appl. Technol.*, 2018.
- [98] C. Harteveld and S. Sutherland, "The goal of scoring: Exploring the role of game performance in educational games," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2015, pp. 2235–2244.
- [99] D. Jackson, S. Kim, C. Lee, Y. Choi, and J. Song, "Simulating Déjà Vu: What happens to game performance when controlling for situational features?," *Comput. Hum. Behav.*, vol. 55, pp. 796–803, 2016.
- [100] B. Lehman, D. Hebert, G. Jackson, and L. Grace, "Affect and experience: Case studies in games and test-taking," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2017, pp. 917–924.
- [101] J. Hao and R. Mislavy, "The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design," *ETS Res. Rep. Ser.*, vol. 2018, no. 1, pp. 1–16, 2018.
- [102] H.-H. Yu, J.-K. Yang, H.-W. Chen, T.-F. Yu, and H.-T. Hou, "Jingnan campaign© - Using game-based assessment with the mechanism of strategy games for history teaching: System development and learning evaluation," in *Proc. 4th Int. Congr. Adv. Appl. Inform.*, 2015, pp. 727–728.
- [103] S.-C. Shih, B.-C. Kuo, and S.-J. Lee, "An online game-based computational estimation assessment combining cognitive diagnostic model and strategy analysis," *Educ. Psychol.*, vol. 39, no. 10, pp. 1255–1277, 2019.
- [104] E. Ibrayamova and G. Stefanov, "Developing and implementing a labyrinth game for self-assessment," in *Proc. 21st Int. Conf. Comput. Syst. Technol.*, 2020, pp. 106–110.
- [105] D. Sanchez and M. Langer, "Video game pursuit (VGPU) scale development: Designing and validating a scale with implications for game-based learning and assessment," *Simul. Gaming*, vol. 51, no. 1, pp. 55–86, 2020.
- [106] M. Ponticorvo, F. Ferrara, R. Di Fuccio, A. Di Ferdinando, and O. Miglino, "SNIFF: A game-based assessment and training tool for the sense of smell," *Adv. Intell. Syst. Comput.*, vol. 617, pp. 126–133, 2017.
- [107] A. S. Robberts and L. Van Ryneveld, "Design principles for introducing 21st century skills by means of game-based learning," *Ind. Higher Educ.*, vol. 36, pp. 824–834, 2022.
- [108] M. J. Dondlinger, "Educational video game design: A review of the literature," *J. Appl. Educ. Technol.*, vol. 4, no. 1, pp. 21–31, 2007.
- [109] J. P. Gee, "Are video games good for learning?," *Nordic J. Digit. Lit.*, vol. 1, no. 3, pp. 172–183, 2006.
- [110] M. Akcaoglu and M. J. Koehler, "Cognitive outcomes from the game-design and learning (GDL) after-school program," *Comput. Educ.*, vol. 75, pp. 72–81, 2014.
- [111] M. H. Hussein, S. H. Ow, M. M. Elais, and E. O. Jensen, "Digital game-based learning in K-12 mathematics education: A systematic literature review," *Educ. Inf. Technol.*, vol. 3, pp. 1–33, 2021.
- [112] C. Ware, *Information Visualization: Perception for Design*. 20Burlington, MA, USA: Morgan Kaufmann, 2019.
- [113] K. Cukier, "A special report on managing information," *Economist*, vol. 394, no. 8671, pp. 3–18, 2010.
- [114] J. Chin, R. Dukes, and W. Gamson, "Assessment in simulation and gaming: A review of the last 40 years," *Simul. Gaming*, vol. 40, no. 4, pp. 553–568, 2009.
- [115] R. J. Mislavy and G. D. Haertel, "Implications of evidence-centered design for educational testing," *Educ. Meas., Issues Pract.*, vol. 25, no. 4, pp. 6–20, 2006.
- [116] V. J. Shute, "Stealth assessment in computer-based games to support learning," *Comput. Games Instruct.*, vol. 55, no. 2, pp. 503–524, 2011.

- [117] C. I. Johnson and R. E. Mayer, "Applying the self-explanation principle to multimedia learning in a computer-based game-like environment," *Comput. Hum. Behav.*, vol. 26, no. 6, pp. 1246–1252, 2010.
- [118] A. All, E. P. N. Castellar, and J. Van Looy, "Measuring effectiveness in digital game-based learning: A methodological review," *Int. J. Serious Games*, vol. 1, no. 2, pp. 3–20, 2014.
- [119] F. Fidler and J. Wilcox, "Reproducibility of scientific results," *Stanford Encyclopedia Philosophy*, 2018.
- [120] T. van der Zee and J. Reich, "Open education science," *AERA Open*, vol. 4, no. 3, 2018, Art. no. 2332858418787466.
- [121] S. Buck, "Solving reproducibility," *Science*, vol. 348, no. 6242, pp. 1403–1403, 2015.
- [122] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–52.
- [123] G. Petri and C. G. von Wangenheim, "How games for computing education are evaluated? A systematic literature review," *Comput. Educ.*, vol. 107, pp. 68–90, 2017.
- [124] M. J. Gomez, J. A. Ruipérez-Valiente, P. A. Martínez, and Y. J. Kim, "Applying learning analytics to detect sequences of actions and common errors in a geometry game," *Sensors*, vol. 21, no. 4, 2021, Art. no. 1025.
- [125] M.-C. Li and C.-C. Tsai, "Game-based learning in science education: A review of relevant research," *J. Sci. Educ. Technol.*, vol. 22, no. 6, pp. 877–898, 2013.
- [126] M. Freire, á. Serrano-Laguna, B. Manero, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, "Game learning analytics: Learning analytics for serious games," in *Learning, Design, and Technology*. Berlin, Germany: Springer, 2016, pp. 1–29.
- [127] C. S. Loh, Y. Sheng, and D. Ifenthaler, "Serious games analytics: Theoretical framework," in *Serious Games Analytics*. Berlin, Germany: Springer, 2015, pp. 3–29.



Manuel J. Gomez received the B.Sc. degree with a focus on applied computing and data science from the University of Murcia, in 2020 and the M.Sc. degree in big data from the University of Murcia, in 2021. He is currently working towards the Ph.D. degree in computer science with the University of Murcia, Murcia, Spain.

He is a member of the CyberDataLab at the University of Murcia. His research interests include data mining, educational technology, game-based assessment, and natural language processing.



José A. Ruipérez-Valiente (Senior Member, IEEE) received the B.Eng. degree in telecommunications from Universidad Católica de San Antonio de Murcia, Murcia, Spain, in 2011 and the M.Eng. degree in telecommunications and the M.Sc. and Ph.D. degrees in telematics, all from the Universidad Carlos III of Madrid, Madrid, Spain, in 2013, 2014, and 2017, respectively, while conducting research with institute IMDEA Networks in the area of learning analytics and educational data mining.


He was a Postdoctoral Associate with MIT. He has received more than 20 academic/research awards and fellowships, has authored or coauthored more than 100 scientific publications in high impact venues, and participated in more than 18 funded projects. He is currently an assistant professor of computer science and artificial intelligence with the University of Murcia, Murcia.



Félix J. García Clemente received the M.Sc. and Ph.D. degrees in computer science from the University of Murcia (UMU), Murcia, Spain, in 2002 and 2006, respectively.

He is an associate professor with the Department of Computer Engineering, UMU. His teaching includes courses in computer networks, network management, ubiquitous computing, and mobile device programming. His major research interests focus on cybersecurity, distributed management of networks and services, and interaction systems.

2 Towards Semantic Interoperability

| | |
|--|----------------------------------|
| Title Developing and Validating Interoperable Ontology-driven Game-Based Assessments | |
| Authors <u>Manuel J. Gomez</u> ¹ , José A. Ruipérez-Valiente ¹ , Félix J. García Clemente ¹ ¹ <i>Department of Information and Communications Engineering, University of Murcia, Spain</i> | |
| Publication details | |
| Journal | Expert Systems with Applications |
| Publisher | IEEE |
| Volume | 248 |
| Number | - |
| Pages | 123370 |
| Year | 2024 |
| JIF | 7.5 |
| Rank | Q1 |
| Status | Published |
| DOI | 10.1016/j.eswa.2024.123370 |
| Abstract <p>Video games have assumed an important place in our daily lives. This has led to an increasing interest on the use of games for non-entertainment purposes, introducing the concept of Serious Games (SGs). In particular, SGs are being explored because of their potential to provide reliable assessments, but also because they can measure competences that would be difficult to measure using traditional forms of assessment. However, one of the key issues is that assessment machinery has to be designed specifically for each game, increasing the time and effort when designing and implementing Game-Based Assessments (GBAs). In this research, we introduce a novel approach to develop interoperable GBAs by: (1) designing and creating an ontology that can standardize the GBA area; (2) conducting a validation study on literature metrics to replicate them and designing novel metrics using data from different SGs; (3) conducting a case study that illustrates how our approach can be used in a real life scenario with real data. Our results confirm that the designed ontology can be used to effectively perform GBAs, along with the metrics replicated and designed in the system. We expect our work to solve the current limitations regarding GBA interoperability, thus allowing the deployment of Game-Based Assessments as a Service (GBAaaS).</p> | |
|  | |



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Developing and validating interoperable ontology-driven game-based assessments

Manuel J. Gomez^{*}, José A. Ruipérez-Valiente, Félix J. García Clemente

University of Murcia, Calle Campus Universitario, 30100, Murcia, Spain

ARTICLE INFO

Keywords:

Serious games
Game-based assessment
Interoperability
Ontologies
Data mining

ABSTRACT

Video games have assumed an important place in our daily lives. This has led to an increasing interest on the use of games for non-entertainment purposes, introducing the concept of Serious Games (SGs). In particular, SGs are being explored because of their potential to provide reliable assessments, but also because they can measure competences that would be difficult to measure using traditional forms of assessment. However, one of the key issues is that assessment machinery has to be designed specifically for each game, increasing the time and effort when designing and implementing Game-Based Assessments (GBAs). In this research, we introduce a novel approach to develop interoperable GBAs by: (1) designing and creating an ontology that can standardize the GBA area; (2) conducting a validation study on literature metrics to replicate them and designing novel metrics using data from different SGs; (3) conducting a case study that illustrates how our approach can be used in a real life scenario with real data. Our results confirm that the designed ontology can be used to effectively perform GBAs, along with the metrics replicated and designed in the system. We expect our work to solve the current limitations regarding GBA interoperability, thus allowing the deployment of Game-Based Assessments as a Service (GBaaS).

1. Introduction

Nowadays, technology plays a very important role in our life, making our work much easier and less time consuming. One of the most prominent examples of technology is the use of digital games for learning (De Freitas, 2006). In recent years, video games have assumed an important place in the lives of children and adolescents, impacting on various aspects of everyday life such as our consumption, communities, and identity formation (Daniel & Garry, 2018; Gros, 2007). In fact, two thirds of adults and three quarters of kids under 18 play video games weekly, and during the pandemic, 71% of parents saw video games as a much-needed break for their children (ESA, 2021). While video games are usually associated with entertainment and leisure, they have recently emerged as powerful tools for learning and skills development. This has generated an increasing interest on the use of games in non-entertainment contexts during the last decade (Susi, Johannesson, & Backlund, 2007). Specifically, the concept of Serious Games (SGs) was first coined by Abt (1987), and probably the most common definition is: “games that do not have entertainment, enjoyment, or fun as their primary purpose” (Laamarti, Eid, & El Saddik, 2014). SGs are mainly used in education; however, they are also used in many other domains (Laamarti et al., 2014), including

training, well-being, advertisement, interpersonal communication, or assessment, among others.

SGs are increasingly being explored for use as assessment tools in broad domains, in particular for their potential to provide more valid assessments compared to traditional assessment approaches, also providing more meaningful and authentic contexts for assessments through interactive immersive environments (Kato & de Klerk, 2017). Specifically, Game-Based Assessment (GBA) is a specific application of SGs, referring to a type of assessment that uses players' interactions with the game as a source of evidence to make meaningful inferences to reveal knowledge, skills, and attributes of users and students that are “invisible” or hard to detect when assessed with more traditional assessment methods (de Klerk & Kato, 2017; Gomez, Ruipérez-Valiente, & Clemente, 2022). However, some limitations are still present, hindering the use of GBAs in real world environments. For example, little is known as to what degree of design complexity is required for meaningful learning to occur, and many games are simple designs targeting low level literacy and providing drill and practice methods (Qian & Clark, 2016). Moreover, there is a lack of sound empirical evidence on the effectiveness of GBAs due to different outcome measures for assessing effectiveness, varying methods of data collection and inconclusive or difficult to interpret results (All, Castellar, & Van Looy, 2014). Usually,

^{*} Corresponding author.

E-mail addresses: manueljesus.gomez@um.es (M.J. Gomez), jruiperez@um.es (J.A. Ruipérez-Valiente), fgarcia@um.es (F.J.G. Clemente).

<https://doi.org/10.1016/j.eswa.2024.123370>

Received 30 August 2022; Received in revised form 16 November 2023; Accepted 28 January 2024

Available online 1 February 2024

0957-4174/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

data collected from SGs are totally different between them, and there is no interoperability between different data sets (Serrano-Laguna et al., 2017). Thus, GBA machinery (including metrics, dashboards, and other type of analytics) are usually designed for each game separately, which increases the time and effort needed building each model from scratch. Since this is one of the key issues open in the area, there is an urgent need to work on data interoperability in order to re-use assessment models and machinery from one game to another.

Previous literature has proposed standard data formats, trying to address these data heterogeneity issues. One example is the work in Serrano-Laguna et al. (2017), which proposed Experience API (xAPI), an interaction model to track user activities within learning environments. However, this and other similar approaches are not supported by most SGs. Other technologies that can help us to address data interoperability are Semantic Web technologies. In particular, ontologies capture knowledge about a certain domain and offer an explicit common conceptualization on it (Fathy, Gad, & Badr, 2019). Ontologies are content theories about the classes of individuals, properties of individuals, and relations between individuals that are possible in a specified domain of knowledge (Panov, Džeroski, & Soldatova, 2008). Although the use of games for assessment has enjoyed great popularity and success in recent years, there is a distinct lack of a generally accepted framework that would describe and unify the area of GBAs.

In this research, we introduce a new approach that uses ontologies in order to develop interoperable GBAs, using in-game metrics that are automatically computed processing the provided data. With that purpose in mind, we propose the design of a new GBA ontology, and we validate it using previous GBA metrics present in literature, as well as new metrics proposed to show the potential that this approach has to perform interoperable GBAs effectively. To validate our ontology, we use data from many different SGs, and we present a case study to demonstrate how this approach could be used in real world situations and environments. Specifically, we have the following objectives:

1. To develop an ontology that can standardize the GBA area, creating a common knowledge model that can integrate the log events from a wide variety of games into our ontology model.
2. To conduct a validation study on previous metrics in literature, as well as the design and implementation of novel metrics using data from different SGs, demonstrating and validating that our approach can effectively perform interoperable GBAs.
3. To conduct a case study that illustrates how our ontology, along with previous and newly developed metrics, can be used in a real scenario.

The rest of the paper is structured as follows: Section 2 reviews background literature on SGs and assessment, GBA metrics and models, and ontologies. Section 3 describes the methodology followed to conduct the research, as well as the games and the data collection used. Next, Section 4 present the results (including the ontology developed), a set of interoperable in-game metrics, and finally the case study conducted. Then, we finalize the paper with discussion in Section 5 and conclusions and future work in Section 6.

2. Related work

In this section we present a review of the literature in some areas which are related to our work: in Section 2.1 we present literature related to serious games and assessment; in Section 2.2 we review some GBA and metrics background and previous studies; and finally in Section 2.3 we review literature related to ontologies and their use in interoperable environments.

2.1. Serious games and GBA

In recent years, many new ways of teaching academic and professional skills to children and adults have been tested using multimedia technologies in the form of software products, educational computer games or video games (Girard, Ecalle, & Magnan, 2013). Reyes-Chua and Lidawan (2019) reported a summary of benefits of using games for learning, including increased learner motivation, reduced learning anxiety, or encouraged creativity and cooperation, among others. In addition, games can promote user engagement through fantasy, interactivity, and non-linear narratives in visual and multisensorial environments that take advantage of advancing technologies (Kato & de Klerk, 2017).

SGs are being used in many different contexts: in education, interest in educational games is continuously growing, but their integration in teaching is still somewhat unexplored area of study (Kangas, Koskinen, & Krokfors, 2017): for instance, some studies have reported difficulties when obtaining an optimal game design, since it is an interdisciplinary task, requiring the contribution of experts from many different areas such as graphic design, product design, programming, or animation (Theodosiou & Karasavvidis, 2015). Moreover, the strict educational system and the fact that some teachers refuse the idea of using “toys” in classroom also entails an added challenge when incorporating games in the classroom (Lee, Luchini, Michael, Norris, & Soloway, 2004). Furthermore, SGs and GBAs have been promoted for use in employee selection as a potential method to improve the user experience, and the use of games in the workplace is a growing phenomenon with SGs being increasingly used as evaluative tools (al Qallawi & Raghavan, 2022). Regarding healthcare, SGs, particularly adventure and shooter games, already play an important role in education, prevention and rehabilitation (e.g. to enhance health-related physical activity, improve sensory-motor coordination, prevent asthma, change nutrition behavior and alleviate diabetes and prevent smoking or HIV) (Wiemeyer & Kliem, 2012). Concerning employee training, SGs are being used by corporations of all sizes (Larson, 2020) to train, for example, financial indicators (Donovan & Lead, 2012) or call center assistants (Hinton, 2016; Mollick & Werbach, 2015).

Games are being explored in particular for their potential for assessment, providing promising possibilities for more valid and reliable measurement of users’ skills as compared to the traditional methods of assessment like paper-and-pencil tests or performance-based assessments (de Klerk & Kato, 2017). GBAs can provide more detailed and reliable information, and the emerging interest in this field reflects the need for alternative assessment tools to overcome limitations that are present in classic methods (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013). In contrast with traditional methods, digital GBA methods have the following merits: (a) they are fun and can reduce test anxiety; (b) they allow for recording users’ interactions in detail (i.e., via the accumulation of log data generated by keystrokes and mouse clicks); and (c) they can be designed to provide real-time learning supports (Shute & Ventura, 2015). In literature, we can see that previous studies have integrated GBAs in classrooms in order to assess students’ skills or knowledge in many different domains: mathematics (Chiu & Hsieh, 2017), art (Basu et al., 2020), language (Song & Sparks, 2019), or soft skills (Nikolaou, Georgiou, & Kotsasarlidou, 2019) are only a few examples of knowledge domains where researchers have conducted studies using games for educational assessment. Apart from education, we can also find previous literature applying GBA in other different contexts, such as healthcare (Vallejo et al., 2017) or employee selection (Georgiou, Gouras, & Nikolaou, 2019). However, the GBA potential to perform valid assessments is mitigated by the time and effort that designing these types of assessments require. Therefore, a common way to conduct GBAs without going through the entire design and implementation process would alleviate these issues and promote more assessments using SGs. In our work, we introduce an intermediate layer that acts as a common knowledge model in order to be used with different data formats so that GBAs can be performed and visualized by simply adapting the data available.

2.2. Game-based assessment models and metrics

Once we have gathered data from users' interaction with a specific game, how to perform a valid and reliable assessment? Although some game and learning analytics can indeed be used in GBAs, they lack specific metrics and methods that outline their effectiveness. SGs analytics need to provide (actionable) insights that are of values to the stakeholders (Loh, Sheng, & Ifenthaler, 2015). In education, new techniques such as Learning Analytics (LA) are trying to provide insight about the educational processes and improve the common educational scenarios benefiting from data-driven approaches (Alonso-Fernández et al., 2019). Its aim is to understand learners and their environments, and improve the learning process through analysis of the data collected from students' interactions with the learning environment to assess students, predict future events and act consequently to refine educational actions (Alonso-Fernandez, Calvo-Morata, Freire, Martinez-Ortiz, & Fernández-Manjón, 2019; Serrano, Marchiori, del Blanco, Torrente, & Fernández-Manjón, 2012).

LA and other techniques, such as data mining (and educational data mining), can be used to fuel the advancement of games research through leveraging the rich data streams enabled by digital GBAs (Owen & Baker, 2019). These areas are applied to explore models and techniques for making efficient and effective use of these data: capturing, tracking, aggregating, analyzing, and visualizing/utilizing information about users' interactions with learning content and their learning progress (Shoukry, 2020). The use of data from games can be collected while users are playing to analyze not only the impact the game is making (in their learning), but also the appropriateness of the game design and its mechanics (Alonso-Fernández et al., 2019).

Gathered data should help get inferences about general traits and abilities of the learner, his general knowledge state, his situation-specific state, his behaviors and his outcomes (Shoukry, Göbel, & Steinmetz, 2014). A frequent approach is to use a set of metrics (or indicators) calculated from users' data. In fact, assessment mechanics help game designers select or design game mechanics that generate useful game metrics (Plass et al., 2013). Many studies have been using metrics to measure students' interaction with educational games, measuring persistence (DiCerbo, 2014) or engagement (Ruipérez-Valiente, Gaydos, Rosenheck, Kim, & Klopfer, 2020), among others. In a survey conducted by Gris and Bengtson (2021) on assessment measures in game-based learning research, 91 studies were analyzed, and results showed that learning aspects are much more assessed than engagement and usability features. Moreover, metrics can be used for other purposes rather than to report users' knowledge. For example, Martínez, Gómez, Ruipérez-Valiente, Pérez, and Kim (2020) developed a series of metrics related to students' activity (e.g., active time, number of different events), but also to the difficulty of different levels in the game, so teachers can adapt their teaching based on these data. Finally, we can highlight the work by Hamdaoui, Khalidi Idrissi, and Bennani (2016), who used in-game metrics to define the students' learning and playing style, but also to adapt gameplay and learning content based on those metrics. Although each environment may have specific metrics some are more common across environments, such as those related to the activity with numbers of events or active time (Ruipérez-Valiente, Gomez, Martínez, & Kim, 2021).

In Liu, Kang, Liu, Zou, and Hodson (2017), authors performed a systematic review on the use of LA for assessment in games, and the results highlighted the promise of using multiple data sources, as well as combining emerging techniques such as visualization and traditional analyses such as statistical and qualitative analyses. When done correctly, visualization can reveal information otherwise unobtainable through traditional statistical analysis. Information visualization is a field of study in its own right and increasingly includes new approaches to visualize spatial and temporal data for reporting and communication purposes (Loh & Sheng, 2015). In recent years, several dashboard applications have been developed to support learning or teaching.

Such dashboards provide graphical representations of the current and historical state of a learner or a course to enable flexible decision making using visual elements (Podgorelec & Kuhar, 2011; Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). They allow the data to be processed so that they can be visualized in a way that enables the teacher or learner rather than the software to make sense of them, converting the abstract and complex to the concrete and visible by amplifying human cognition (Card, 1999; Duval, 2011).

These metrics and visualization systems present in literature are designed specifically for each game and study, which creates interoperability issues. An objective in our research is to review previous GBA metrics, as well as designing novel interoperable metrics to demonstrate that is possible to perform GBA with different SGs and data formats. In addition, our case study also includes an example of how to use visualizations to graphically represent all these interoperable metrics.

2.3. Interoperability, standards and ontologies in games

While the use of SGs has extended rapidly to a variety of domains, their design, development and later analyses of results remains a challenging individual process for both developers and teachers/trainers (Stănescu, Stefan, Kravcik, Lim, & Bidarra, 2013). Interoperability is a key requirement for organizations regardless of the field they operate. People, organizations and software systems must communicate between and among themselves. However, due to different needs and background contexts, there can be widely varying viewpoints and assumptions regarding what is essentially the same subject matter (Uschold & Gruninger, 1996). The way to address these problems is to reduce or eliminate conceptual and terminological confusion and come to a shared understanding. Previous studies have proposed approaches trying to standardize analytics into games. Alonso-Fernandez, Calvo, Freire, Martinez-Ortiz, and Fernandez-Manjon (2017) proposed an interaction model (xAPI) that can be used to describe streams composed of actors performing with actions in a specific context. Each xAPI statement represents a learning activity and has three main attributes: an actor, a verb and an object: who did what action, with a target of the action and certain additional attributes. Moreover, Perez-Colado, Rotaru, Freire, Martinez-Ortiz, and Fernandez-Manjon (2018) proposed a method that comprises the specificities of location-based games, as an extension of the xAPI standard to support location-based SGs.

Another area that could help us to establish a common model and remove conceptual confusion is the area of ontologies. Ontologies are often defined as a set of concepts, their definitions and their inter-relationships about certain domain (Uschold, 1996). In computer science, the concept "ontology" is interpreted in many different ways and concrete ontologies can vary in several dimensions, such as degree of formality, authoritativeness or quality (Happel & Seedorf, 2006). Researchers in many areas have all recognized the need for ontologies to clearly define specialized vocabularies for these domains (McDaniel & Storey, 2019), and nowadays we can see ontology-based applications in areas as diverse as customer support and engineering of cars (Staab & Studer, 2010). Several approaches have been proposed for developing ontologies (Corcho, Fernández-López, & Gómez-Pérez, 2003): following a bottom-up strategy, on the basis of an application Knowledge Base (KB); to reuse large ontologies to build domain specific ontologies and KBs; collaborative construction (agreeing new pieces of knowledge with the rest of the knowledge architecture, which has been previously agreed)... However, it is not usually necessary to implement ontologies manually, as most of the available ontology tools are able to generate ontologies in many different ontology languages. Although we identified some studies that have proposed approaches trying to standardize the use of games for assessment (e.g., Said, Cheniti-Belcadhi, & El Khayat, 2019 with an ontology for personalization in serious games for assessment, or Tang & Hanneghan, 2011 with an ontology for serious game design), we did not found any study trying to standardize the GBA area using an ontology-based approach.

As a result of students' interaction with games, large repositories of data are generated. When it comes to the relationship between ontologies and this vast amount of data, data is usually stored in the computer main memory; thus, some problems exist when manipulating a large amount of ontology-based data (Dehainsala, Pierra, & Bellatreche, 2007). Trying to solve these issues, some studies have incorporated the use of big data technologies when managing ontology-based data. Some examples are the use of MongoDB (a NOSQL database) and modular ontologies (Abbes & Gargouri, 2016), or the use of Spark and Flink (two data processing frameworks) for the construction of an engine for scalable processing of large-scale RDF data (Lehmann et al., 2017). In this research, we plan to go beyond literature, trying to standardize the GBA area defining a common knowledge model. With this purpose in mind, we combine the use of ontologies and GBA metrics in order to address the interoperability issue in this particular area.

3. Methodology

In this section, we present the selection of games and metrics that have been used for validation, as well as the construction method that we have followed to build our ontology. In addition, we introduce the framework where our case study and validation has been conducted.

3.1. Serious games selection

Since we wanted to test our ontology in a real scenario, we selected a set of different SGs within different knowledge domains to test the interoperability of our approach. Field Day (Field Day Lab, 2022) is a research lab based at the Wisconsin Center for Education Research at the University of Wisconsin - Madison. Field Day designs learning games that bring contemporary research to the public, making their game data available to the public. Exploring this open game data (Gagnon & Swanson, 2023), we made a selection of SGs that we use in our study:

- *Magnet hunt*: a game where learners have to use magnetic waves to find a set of magnets hidden throughout the yard. This game addressed the topic of magnetism, a class of physical phenomena that are mediated by magnetic fields. Moreover, it also addresses other topics like forces and interactions, magnetic poles, and magnetic fields.
- *Wave combinator*: a game where students learn how waves interact using a mysterious wave combinator found in the yard. This game addresses wave properties, amplitude, offset, wavelength, frequency, and more topics.
- *Crystal cave*: a game where students learn all about crystal molecules and dig up some sweet crystals for their collection in the museum. This game addresses crystals, geometric arrangement, molecular charges, and molecular stability topics.
- *Wind simulator*: using this simulator, students learn how wind travels from high to low pressure systems, but moves in a spiral due to the coriolis effect. The game addresses concepts related to earth's systems, air masses and weather conditions, and weather patterns.
- *Antibiotic resistance*: Playing this game, learners will acquire knowledge about heredity, inheritance, variation of traits, mutation, genes, and antibiotics.
- *Earthquake*: in this game, students learn about real earthquakes, with concepts such as earth's materials and systems, scale proportion and quantity, S waves, P waves and triangulation.
- *Nitrogen cycle*: learners have to figure out how nitrogen atoms move around the world to win the game. In this process, they will learn concepts such as the nitrogen cycle, bacteria digestion, plant death, plant assimilation, or herbivorism.

Table 1

Dataset sizes.

| Game | Size (MB) | # of events | # of triples |
|----------------|-----------|-------------|--------------|
| Magnet | 50.8 | 100,000 | 1,788,937 |
| Waves | 63.0 | 100,000 | 1,741,566 |
| Crystal | 67.9 | 100,000 | 1,724,822 |
| Wind | 64.1 | 100,000 | 1,646,667 |
| Bacteria | 57.1 | 100,000 | 1,711,218 |
| Earthquake | 44.6 | 100,000 | 1,528,762 |
| Nitrogen Cycle | 55.2 | 100,000 | 1,708,911 |
| Carbon Cycle | 52.2 | 100,000 | 1,719,099 |
| Shadowspect | 44.0 | 100,000 | 1,682,333 |
| Lakeland | 71.3 | 100,000 | 1,821,143 |

- *Carbon cycle*: learners have to figure out how carbon atoms move around the world, mastering the carbon cycle in order to collect enough carbons to beat the final opponent. Learners will acquire competences related to the carbon cycle, systems and systems models, cycle of matter, and energy transfer in ecosystems.
- *Shadowspect*: a geometry game designed explicitly as a formative assessment tool to measure math content standards (e.g. visualize relationships between 2D and 3D objects). It aims to provide metrics related to geometry content and other behavioral and cognitive constructs.
- *Lakeland*: in this strategic building game, learners decide to build a new town called Lakeland, explore the dynamics of the nutrient system and recognizing the impact humans have on the world. The game addresses the Next Generation Science Standards essential practice of Modeling alongside the cross-cutting concepts of patterns, cause and effect, and systems and system models.

All these games (and their data) can be checked and played in <https://felddaylab.wisc.edu/opengamedata/>. As an example, we can see two screenshots of two different SGs from this research lab in Fig. 1. Datasets are usually in TSV format, and, although each dataset has some specific columns, most of them have columns in common that are essential in game log data, such as the "session id", "game id", "timestamp", "game version", or "user id". Data format employed in these datasets has been considered as the base format that is used later as data input in our experiments.

3.2. Data collection

To test the interoperability of our approach, we used ten data sets from the SGs presented in Section 3.1. The size of each dataset is presented in Table 1. As we can see, each one of the datasets contains a total of 100,000 game events derived from real players' interaction with the different games. The number of triples in our experiments varies from 1,528,762 to 1,821,143, depending on each game.

3.3. Metric selection

In this section, we describe the process followed to select and develop the set of metrics that are used to perform GBA and test our approach's capabilities. We can divide these metrics into two different subsets: *literature* metrics, and *author proposed* metrics. First, we wanted to demonstrate that our framework can replicate any metric present in literature. With that purpose in mind, we used the selection of papers of our previous systematic review on the GBA area (Gomez et al., 2022), carefully reviewing each paper and selecting the metrics described. Since our objective was to select metrics in literature, we excluded calculations over data that included Machine Learning (ML), Deep Learning (DL) and similar models/algorithms. Second, we wanted to demonstrate the possibilities that our approach provides creating a set of metrics that go beyond the state of the art, introducing new ways to perform GBA using ontology-based data and SparQL queries. To validate that our ontology-driven metrics were correct, we implemented some of the metrics directly using the collected primary data to check that results were the same using both approaches.

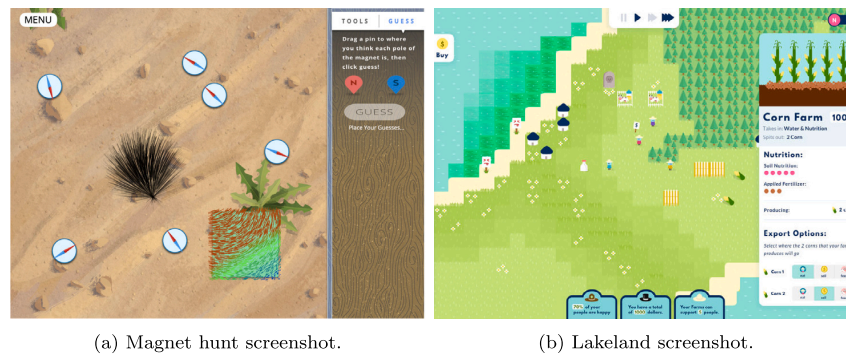


Fig. 1. Screenshots of two of the SGs selected.

3.4. Ontology development process

In general, methodologies give you set of guidelines of how you should carry out the activities identified in the ontology development process, what kinds of techniques are the most appropriate in each activity and what products each one produces (Fernández-López, Gómez-Pérez, & Juristo, 1997). When constructing an ontology, there are two different methods that we can follow, one is to build a new domain ontology directly, and the other is to expand an existing domain ontology. When choosing ontology construction methods, we should choose the most appropriate method according to the actual situation, or even integrate the advantages of various methods (Sun, Hu, Li, & Wu, 2020). To the best of our knowledge, there is not an existing ontology meeting our requirements and that could be used to expand it, so we decided to build our own ontology from scratch.

For building our ontology, we decided to use Methontology (Fernández-López et al., 1997), a structured method designed to build ontologies from scratch, reusing others as they are, or by a process of re-engineering them. Methontology was stated as the most mature approach for building ontologies, being recommended by the Foundation for Intelligent Physical Agents (FIPA) for the ontology construction task (Corcho et al., 2003). In Fig. 2 we can see the complete development process that we have followed to build and validate our ontology, which is an adaptation of the Methontology original methodology and the one proposed by Olszewska et al. (2020), which is also based on Methontology. Next, we explain each step in detail.

3.4.1. Pre-development activities

The pre-development activities include planification, the environment study, and the knowledge acquisition:

- **Planification and environment study:** this is the first phase of the process. These activities consisted in identifying the problem to be solved with the ontology, the applications where the ontology will be validated and integrated, and verifying that the ontology was possible to build, also considering the limitations of the project (Olszewska et al., 2020).
- **Knowledge acquisition:** was thought as an independent activity in the development process. However, it can be conducted simultaneously with other activities, as most of the acquisition is done with the requirements specification phase. It deals with the acquisition of knowledge from experts or other sources, that can include books, figures, tables, brainstorming techniques, or even other ontologies, among others.

3.4.2. Development states

The development states constitute the main core of the methodology, consisting of:

- **Specification:** the goal of the specification phase is to produce an informal, semi-formal, or formal ontology specification document written in natural language, including (a) the purpose of the ontology, (b) the level of formality of the implemented ontology, and (c) the scope, using a set of intermediate representations or competency questions.
- **Conceptualization:** captures the relevant domain knowledge building a conceptual model describing the problem and its solution. The core concept dictionary must meet the requirements of being unambiguous while covering the entire domain. Moreover, we must take into account the concepts' relationships and attributes. In this vein, we can build a set of intermediate representations such as a glossary of terms, a verb dictionary, or tables of rules and formulas, if needed.
- **Integration:** it explores the use of other ontologies to speed up the construction of your ontology, reusing certain terms or definitions.
- **Implementation:** in this phase, the formal models built previously are converted into a computable model. As the ontology development environment, we decided to use *Web-Protégé*, a web-based lightweight ontology editor which combines the Google Web Toolkit for the user interface, and Protégé for supporting ontology services. It is open source, and also provides collaborative features to facilitate discussions and annotations between different contributors (Tudorache, Vendetti, & Noy, 2008).
- **Validation in a real scenario:** since the main goal of our study was to perform interoperable GBAs in a real context, this phase includes the validation of the ontology using a real dataset, collected as a result of the users' interaction with different games that have been used in real life. In this state, we validate the ontology by trying to represent the information contained in the dataset and checking if the initial objectives defined are accomplished.
- **Formal evaluation:** this stage means to carry out a technical judgment of the ontology and their software environment with respect to a frame of reference. The formal evaluation includes (1) *Verification* (i.e., the technical process that guarantees the correctness of an ontology) and (2) *Validation* (i.e., guarantee that the ontology and the software environment correspond to the system that they are supposed to represent). To perform the *verification*, we used OOPS! (Ontology Pitfall Scanner!) (Poveda-Villalón, Gómez-Pérez, & Suárez-Figueroa, 2014), a tool for detecting pitfalls in ontologies, which operates independently of any ontology development platform and is available online. For example, OOPS! warns you when the domain or range of a relationship is defined as the intersection of two or more classes, or when a cycle between two classes in the hierarchy is included in the ontology, which could lead to reasoning problems.

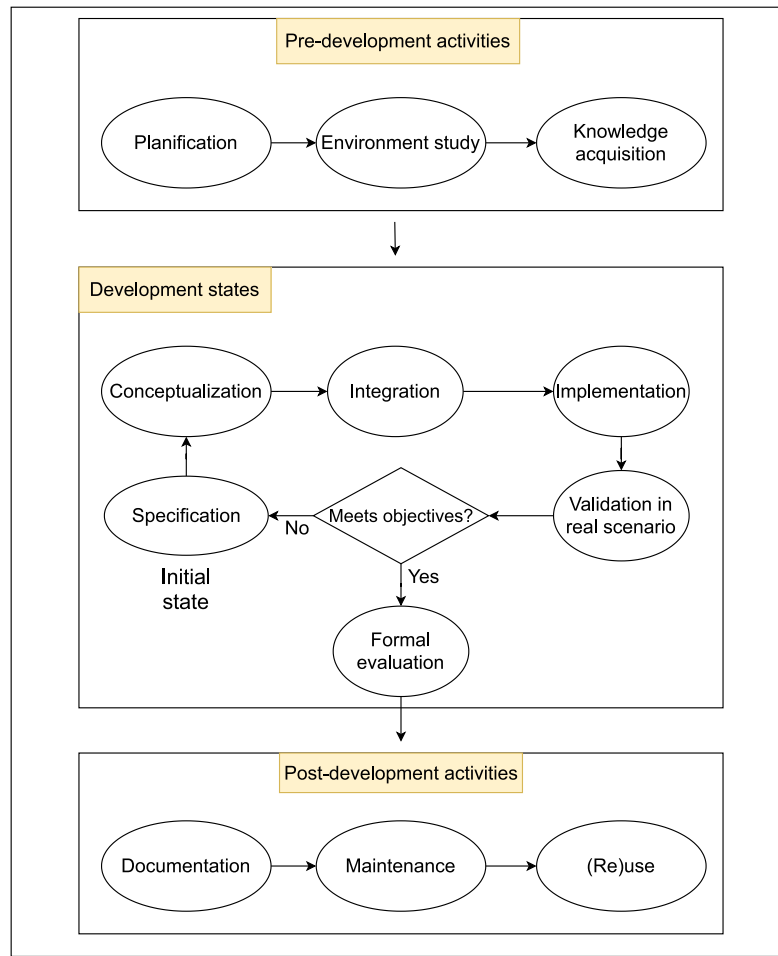


Fig. 2. Ontology development process.

Note that this is an iterative process: after the ontology is initially constructed, it can be evaluated and improved. If in the “Validation in real scenario” stage we discover any type of inconsistency or any objective that is being not satisfied, we then go again to the initial development phases to fix any problem and keep improving the ontology.

3.4.3. Post-development activities

This part of the development is done only when the previous phases have been finished. The post-development activities include:

- **Documentation:** this phase gather all the documents produced during the previous stages in order to create an appropriate ontology documentation.
- **Maintenance:** updates the ontology creating/removing concepts or relationships, allowing the ontology to evolve based on new applications that it could have.
- **(Re)use:** it considers the use of the ontology for the original purpose, but also the reuse of the developed ontology in other ontologies and/or applications.

Although these activities are only executed once, the maintenance and (re)use phases can be repeated if necessary after the ontology development has been accomplished.

3.5. Framework to support interoperable game-based assessments as a service

Once the ontology has been created and validated, we need a powerful tool capable of integrating our ontology and using ontology-based data to develop useful metrics. Consequently, we decided to use a framework developed to support interoperable Game-based Assessments as a Service (GBAaaS) (Gomez, Ruipérez-Valiente, & García Clemente, 2023). The complete framework’s architecture is shown in Fig. 3.

As we can see, the framework uses standard formats for data input (e.g., CSV, TSV), transforming these data into ontology-based data in Resource Description Framework (RDF)/XML format. RDF is a general-purpose language for representing data and metadata on the web, and it is supported by its own query language SparQL, enabling the extraction and transformation of RDF data (Gandon, Bottollier, Corby, & Durville, 2007). In the next step, SANSa framework (Lehmann et al., 2017) is used as a base to process these ontology data and infer new information from it, as well as perform queries over the inferred data. All metrics designed and developed have been implemented in form of SparQL queries. Then, thanks to the metric output module, query results can be exported using several formats, such as plain text or CSV. In addition, the framework also provides a REST API module, allowing to use it as an online service. Finally, the authentication and authorization module

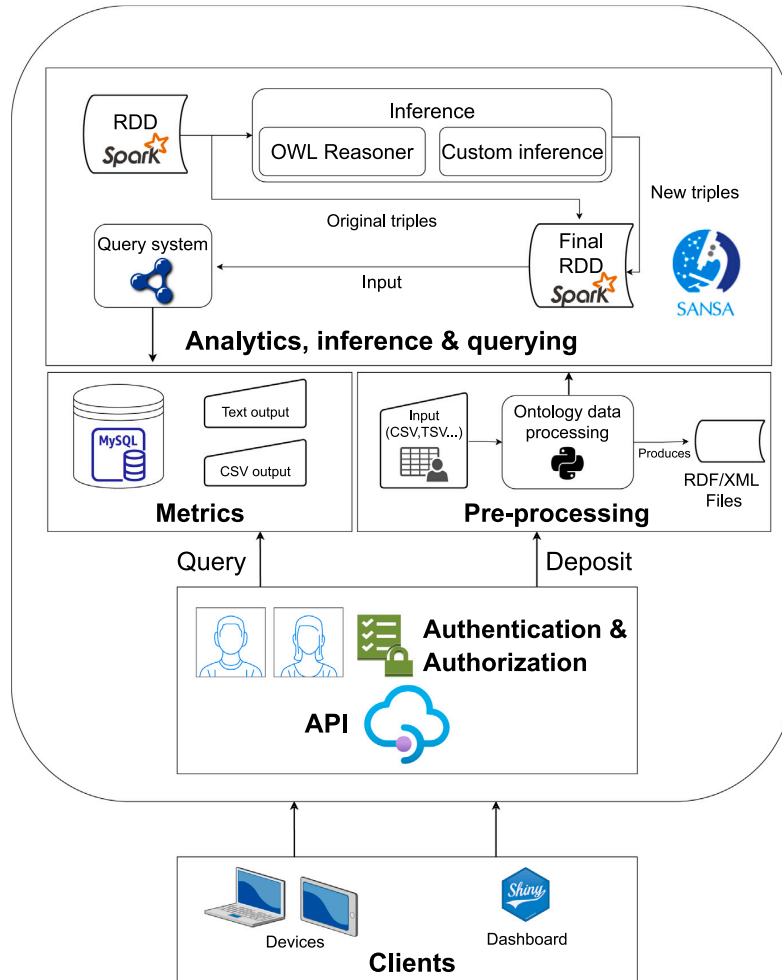


Fig. 3. GBA framework's architecture.

manage the different roles and authorizations in the system, making sure that clients making petitions through the API module are properly logged in and have the right permissions to operate. A comprehensive description of this framework and its components can be consulted in our previous work (Gomez et al., 2023).

4. Results

4.1. Proposed ontology

The ontology has been developed following the methodology described in Section 3.4. We only include the results derived from the phases “Specification”, “Conceptualization” and “Implementation”, since the rest of activities and phases do not have a specific output.

4.1.1. Specification

This activity produces a specification document as an output. As stated in our methodology, the specification document must address the most vital questions related to the domain we are interested in, the ontology purpose, and its scope. In addition, we have proposed a set of Competency Questions (CQs), which are the questions that are later used at the evaluation phase to assure that the ontology is appropriate for the purpose originally thought. In Table 2 we can see the specification document proposed for our ontology.

4.1.2. Conceptualization

This activity aims to capture the relevant knowledge with a set of intermediate representations. For our ontology, we built a core concept dictionary (which includes the main terms and concepts involved in the GBA domain, along with possible synonyms), a table of attributes (including the attributes of each concept), and a binary relation table (which includes the relationships between the different concepts and its cardinality). In Table 3 we can see the core concept dictionary of our ontology, including relevant terms such as “Game”, “Game session”, “Player” or “Learning outcome”. Moreover, in Table 4 we can see two examples of the binary relation table, indicating the source and target concepts and their cardinality.

4.1.3. Implementation

In this step, we prepared and converted the GBA Ontology into machine readable format, using an ontology development editor. As we stated in the methodology, we use *Web-Protégé* and *Protégé* as ontology modeling tools. The classes constructed and their relationships are shown in Fig. 4.

As we can see, the final ontology model includes the core concepts previously identified in form of classes, as well as a set of relationships that aim to represent the links between the different classes. For example, we see that a user can have a relationship with certain user

Table 2
The ontology requirements specification document.

| Specification Document | |
|------------------------|--|
| Domain | Game-based Assessment. |
| Date | Nov, 9th 2021. |
| Conceptualized by | Research author. |
| Purpose | Ontology about game-based assessments to be used in different contexts and with different types of data. The ontology could be used to infer knowledge about the existing data related to users' assessment, creating new information such as the level of the player, or different play styles. |
| Level of formality | Semi-formal. |
| Scope | Users' assessment using data from serious educational games. |
| Competency questions | <ul style="list-style-type: none"> • How to assess or measure that the required learning objective has been achieved? • Which users have a specific play style (e.g., persistence)? • Which levels have been completed by a user in a game? • How much time has a user spent playing different games? • How users have interacted with different games? |

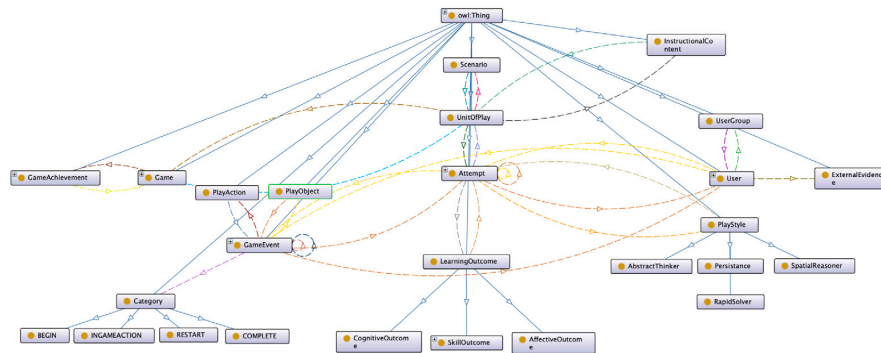


Fig. 4. GBA Ontology classes and relationships visualized via Protégé Ontograp.

Table 3
Core concept dictionary.

| Core concepts and terms | |
|-------------------------|--|
| Concept | Synonym |
| Game | – |
| Game event | Assessment statement |
| Game achievement | – |
| Instructional content | – |
| Unit of play | Level |
| Scenario | Environment |
| Instructional content | – |
| Attempt | – |
| Learning outcome | Capability |
| Game session | – |
| User | Player |
| User group | Player group |
| Play style | Behavior |
| External evidence | – |
| Sources of knowledge | Nouira, Cheniti-Belcadhi, and Braham (2018), Plass, Homer, and Kinzer (2015), Rocha and Zucker (2015), Said et al. (2019), Tang and Hanneghan (2011), Yusoff, Crowder, Gilbert, and Wills (2009) |

Table 4
Binary relation table sample.

| Binary relation table | |
|-----------------------|------------------|
| Relation name | Has |
| Source concept | Game |
| Source cardinality | (1,n) |
| Target concept | Game achievement |
| Target cardinality | (1,n) |
| Relation name | Has |
| Source concept | Game session |
| Source cardinality | (1) |
| Target concept | Attempt |
| Target cardinality | (1,n) |

game(s). In order to view the complete ontology, the reader can create a new account on the *Web-Protégé* web page, and then use the following link: <https://webprotege.stanford.edu/#projects/d0adb6c-e703-4f81-bbfa-9fcd9a974a07/edit/Classes>.

4.2. Metric ontology validation

In this Section we aimed to validate the ontology proposed by using two different groups of metrics: metrics that emerged from previous literature and newly designed metrics. Finally, we also explain how these metrics have been implemented in the framework.

group, but also with game sessions and play styles. These relationships contribute to improve the knowledge and they can be used to perform queries from the ontology with reasoning tasks. For example, they can be used to ask for the game events that certain user has on certain

4.2.1. GBA literature metrics

As we stated previously, we did a careful review of the GBA literature to search and replicate metrics that have been implemented in previous studies. After reviewing and grouping each metric, we created six groups:

- **Activity indicators:** This metric is computed for each game, group and user, and includes the total amount of time spent in the game, the total number of events, and the frequency of events (number of events/total time).
- **Persistence indicators:** This metric is computed for each game, group and user, and includes the total amount spent in units (levels), the number of units completed, and the maximum time spent in a single unit.
- **Event types:** This metric is computed for each user and game, and includes the number of events of each user grouped by event type (e.g., “Complete”, “Retry”, “Interaction”). In addition, this group also includes the interaction level, which is defined as interaction events divided by the sum of the rest of events.
- **User performance:** This metric is computed for each game, group and user, and includes the percentage of success (which is defined as the number of units completed divided by the number of units started), and the maximum unit reached by the player.
- **Levels of activity:** This metric is computed for each game, group and user, and include straightforward metrics to compute based on a feature engineering process, such as the active time, inactive time, number of events, and the number of different types of events.
- **Funnel by user:** This metric is computed for each game, group and user, including the percentage of units that the user has started, the percentage of units that the user has interacted with, and finally the percentage of units that the user has completed. This funnel seeks to provide a quick overview of the current status and progress for each user and game.

These metrics have been implemented in our framework in form of SparQL queries, and they will be used to illustrate how we can use this approach in a real scenario. In addition, the granularity of these metrics can be changed very quickly, being able to aggregate results by whole groups or even the whole game.

4.2.2. Additional metric proposal

Moreover, going beyond this type of operations over data, we designed and implemented a set of metrics that require of more complex calculations, such as standardization, normalization, or more complex methods (e.g., machine learning techniques). Next, we explain each metric in detail:

- **Levels of difficulty:** This metric is computed for each game and unit of play (level), and provides a set of parameters that are related to the difficulty of the different units (Ruipérez-Valiente et al., 2021), namely: *completed_time*, which is computed by dividing the amount of time invested in the game by the number of completed units; *actions_completed*, which is computed by dividing the number of actions by the number of completed units; and *p.abandoned*, computed by dividing the number of started units by the number of completed units. Then, a standardized and normalized measure of the three previous parameters together in a single value is computed, representing the difficulty score of each unit.
- **Persistence:** This metric is computed for each game, user, and unit of play (level). Although there are few indicators on how to calculate persistence, it can be observed that time, both for completed and uncompleted activities, and the number of attempts are essential characteristics for persistence. Following some related works, it can be seen that the rest of the parameters are more linked to the specifications of each scenario where it has

been implemented (Valiente, 2022). In our specific metric, to see if a user has been persistent or not, we consider several metrics: if the unit has been completed or not, active time spent, number of events triggered, and number of attempts. Accordingly, we consider percentiles of each of the parameters considered (time, attempts, events), supposing that the user was persistent if the respective value exceeds the value of 75%. Lastly, for each user, we identify the units in which he has been persistent, and we calculate if the user globally has been persistent or not according to the number of units in which he has been persistent.

- **Play styles:** This metric is computed for each game and user. To identify different play styles that users can have when playing, we perform clustering by using k-means algorithm, which commonly uses a set of continuous variables as input. We use as input the following indicators: total active time, different days played, number of different events triggered, number of interaction events, and completed units. Based on these indicators, we can obtain higher level profiles so that we can analyze each cluster separately and determine different play styles.

For these metrics, the initial calculations required have also been implemented in form of SparQL queries. However, other advanced techniques (such as ML) have been included in the system by developing separated scripts that use the SparQL queries results to perform complex calculations. Moreover, the granularity in these metrics can also be changed easily so that metric results are displayed not only by individual users but also by group or game.

4.2.3. Metric implementation

To implement the metrics presented, we used the framework described in Section 3.5. Thanks to its capabilities, we have integrated the ontology developed along with the different metrics. Once the log data from learners' interaction with SGs has been transformed into ontology-based data, the system uses SparQL (the standard language for querying RDF data) queries to gather information from these data. Therefore, we implemented all the basic metrics in form of queries, including metrics emerged from literature as well as the first stages of the additional metrics proposed. Regarding the additional metrics, more complex calculations were executed using specific scripts integrated into the framework, which also used the results from the SparQL queries. Next, as an example, we show the query implemented for the “Activity indicators” metric:

```
SELECT ?game ?group ?user
  ?totalTime ?totalGameEvents
  ((?totalGameEvents/?totalTime) as ?freqEvents)
WHERE{
  ?user rdf:type m:User.
  {SELECT ?group ?user ?game
    (SUM(?individualTime) as ?totalTime)
    (COUNT(DISTINCT(?ev)) as ?totalGameEvents)
  WHERE{
    ?user rdf:type m:User.
    ?user m:has ?attempt.
    ?user m:hasGroup ?group.
    ?attempt rdf:type m:Attempt.
    ?attempt m:playedInUnit ?unit.
    ?game m:hasUnitOfPlay ?unit.
    ?attempt m:has ?ev.
    ?ev rdf:type m:GameEvent.
    ?ev m:timeBetweenEvents ?individualTime.
  }
  }
  GROUP BY ?group ?user ?game
}
```

As stated in the metric definition, it calculates the total amount of time spent in the game, the total number of events, and the frequency of events of each learner separately. As we can see, the total time is calculated by adding the number of seconds that each event lasts, and this number is aggregated by group, game and user. Moreover, we also see that the frequency is calculated dividing the total number of events by the total time (in seconds) once the results have been aggregated by learner.

4.3. Case study

In this section, using our experiment results, we present a case study exemplifying how our ontology-based approach and the computed metrics could be used in a real life environment. To test the interoperability and usability of our work, we followed previous related work (Díaz et al., 2019; Jayapandian, Zhao, Ewing, Zhang, & Sahoo, 2012; Santos, Dantas, Furtado, Pinheiro, & McGuinness, 2017) and conducted a similar case study.

4.3.1. Dashboard overview

In this use case, we present a visualization dashboard system that uses the data analyzed and transformed into metrics, being consumed via visualizations. This enables instructors to monitor what learners are doing while playing, use these data to adapt their interventions when necessary, or even use these metrics as a part of a formative evaluation. Moreover, this dashboard also allows learners to track their own activity within the games. We have developed the dashboard using Shiny's R framework, and we have deployed it on ShinyApps web server. In our implementation, we have two types of users: on the one side, we have instructors (or teachers) that can use the dashboard to visualize what their learners are doing. Therefore, instructors are able to insert new GBA data, but also to query metrics from games and groups where they are participating. On the other side, we have learners, which are only allowed to query their own metric results. This way, we restrict the access to different groups and games data to ensure the privacy of each user.

Fig. 5 shows the dashboard running on the ShinyApps server. The login page is shown in Fig. 5(a), where the user can log into the system by using a username and a password. Depending on the credentials used, each person has access to different features and different data sets within the dashboard, as each user created in the platform has its own roles and permissions. After login credentials have been initially verified, the user gets access to different functionalities: in Fig. 5(b) we can see the file upload page, where authorized users can upload new GBA assessment data to be processed by the framework and incorporated in form of new metrics data. Moreover, the user can choose between different tabs available in the sidebar, either to upload new data if the current user's role permits it, or to query metric results and see them graphically via visualizations.

Furthermore, Fig. 6 shows two illustrations of how metrics are represented in the system. In our dashboard, we have implemented both group-oriented and individual-oriented visualizations, depending on the granularity of each metric. The dashboard takes advantage of the complete interoperability between games and metrics, as the user can manipulate the selection boxes to filter by different games and groups (as shown in Fig. 6(a)), and also by user if the metric allows it (as shown in Fig. 6(b)). That way, when a game is selected among the available options, the system shows the existing groups for that specific game in the corresponding selection box; once all the selection boxes for that metric tab are filled with a choice, the system queries the necessary information and represents it using interactive visualizations. Next, we present a use case using these visualizations with real data.

4.3.2. Group and student analysis

This use case exemplifies how an instructor can use the dashboard to analyze the global group status, but also to monitor individual learners. This can be very useful to track issues related to the whole group and be aware of the learners' current progress. For this use case, we analyze some of the metrics that have been defined with specific examples, using real data from the game and group that we have selected for the analysis.

In Fig. 7(a) we see the Funnel by user metric visualization, which is based on the metric we defined previously. In this concrete use case, we selected the group "MainGroup" from the game "CRYSTAL". We see that there are 26 learners, with a funnel corresponding to each learner, showing the percentage of units that have been started, interacted, and completed. For example, we can focus on the user with identifier "210604085851886" (third funnel in the first row), which is a learner with a good performance in the game. This learner has started and interacted 100% percent of the units, and completed correctly 89% of them. Then, the instructor could use the "Persistence" visualization (Fig. 7(b)) to have a more detailed perspective of how each learner has interacted with the different units. In this visualization, a pie chart for each user is shown, indicating the percentage of units in which the user has shown different types of behaviors, such as "productive persistence" or "unproductive persistence". Focusing on the same user as before, we see that has shown "no behavior" in 62.5% of the units, "productive persistence" in 25% of the units, and that has been "non persistent" in 12.5% of the units.

Then, the instructor might want to know more about the activity that the learner has had in each unit. With this purpose in mind, an instructor could use the "Levels of activity" visualization shown in Fig. 8(b), adding a higher level of detail regarding learner's interaction. In this visualization, for a given user, we can see the active time, number of events, and number of different events in each unit. For the selected user, we can see that most interacted units have been "CRYSTAL-7" and "CRYSTAL-8", with an active time higher than 200 s, and a number of events of 115 and 66, respectively. Then, to see which have been the more complicated units for the group, the instructor can take a look at the "Levels of difficulty" visualization (Fig. 8(a), in which we can see the parameters defined, and the final difficulty measure calculated for each unit. As we can observe, the most difficult units for these group have been "CRYSTAL-7", and "CRYSTAL-8", which perfectly matches with the interaction patterns previously seen within the selected learner's data.

5. Discussion

Although SGs are being considered as useful tools to perform complex and reliable assessments in broad domains (Kato & de Klerk, 2017; Sliney & Murphy, 2011), the implementation of GBAs features is seen as a very time consuming step (Ifenthaler, Eseryel, & Ge, 2012). This is due to heterogeneity issues, since assessment machinery is usually designed specifically in each different game and context. Previous work has addressed this problem by proposing standard models, such as xAPI, comprising the specificities of analytics in games (Alonso-Fernandez et al., 2017; Perez-Colado et al., 2018). In this research, we try to address the interoperability issue by providing a higher level interoperable approach to perform GBAs. We designed and implemented a new ontology that serves as a common knowledge model, being able to integrate log events from any game into a unified data model. This implies the standardization of a wide area, with games designed with different purposes, based on different knowledge areas, and targeting participants with different characteristics.

Comparing our approach with previous standardized data format approaches, which benefits can bring the use of an ontology to this area? First, ontologies provide an organization and reuse of knowledge, allowing to disambiguate or uniquely identify the meaning of concepts in a given domain (Bürger & Simperl, 2008). Second, ontologies allow

(a) Login page screenshot.

| session_id | app_id | timestamp | event_name | event_data | version | index | group |
|-------------------|----------|-------------------------|------------|--|---------|-------|-----------|
| 21060408415973764 | BACTERIA | 2021-07-01 07:42:18.900 | BEGIN.0 | {'level': 0, 'totalTime': 0, 'server_time': '2021-07-01T02:4 ... | 1 | 0 | MainGroup |
| 21060408415973764 | BACTERIA | 2021-07-01 07:42:31.616 | CUSTOM.2 | {'event_custom': 'BACTERIA_CREATE', 'numberCreatedTotal': 0, ... | 1 | 1 | MainGroup |
| 21060408415973764 | BACTERIA | 2021-07-01 07:42:32.481 | CUSTOM.2 | {'event_custom': 'BACTERIA_CREATE', 'numberCreatedTotal': 1, ... | 1 | 2 | MainGroup |

(b) File upload tab screenshot.

Fig. 5. Screenshots of the dashboard developed.

to take advantage of the richness and complexity of relationships between concepts and entities within a domain. By representing these relationships explicitly, ontologies can provide a better understanding of the GBA area and enable a more sophisticated reasoning. This knowledge representation also provides computational inference, which can help to spot logical inconsistencies to indicate modeling errors. Third, it is reasonable to expect a performance gain in precision and recall when using ontology-based approaches compared with the data mining approaches (Dou, Wang, & Liu, 2015). Furthermore, ontology-based approaches can be combined with Big Data technologies in order to obtain great performance results, as shown in Lehmann et al. (2017). In fact, using the framework mentioned in Section 3.5, we were able to compute data from approximately 39 full classrooms during an entire month in 107.2 min (Gomez et al., 2023).

Alonso-Fernandez et al. (2019) conducted a review regarding applications of data science to game learning analytics data, and noted that most papers did not report the format in which they collected the data, so it is unknown if they were using a standard or relying on their own data formats, which leads to reproducibility and reusability problems. Our ontology works regardless of the data format, which can be easily adapted to be incorporated in the model and used for

further processing. This enables an easy way to use any type of GBA data and make straightforward assessments by simply adapting the data to our model. In addition, authors also noted that sample sizes used in the studies are, in general, quite low, presenting low statistical power and having a reduced chance of detecting actual effects (Petri & von Wangenheim, 2017). Thus, it is important that future research used larger data samples, in order to improve the results' generalization, and also to enable the use of more complex techniques that usually require a big amount of data. Our work also enables the processing of large data samples, since it has been integrated in a framework that uses Big Data technologies (specifically Spark) to process the ontology-based data.

The availability of real-time information about the learners' interaction and behaviors provides a great opportunity to analyze these data during gameplay. The analysis of those actions and the investigation of more complex series of actions and behaviors can provide key insights into ongoing learning processes in these environments (Kim & Iffenthaler, 2019). GBAs aim to convert learner-generated information into actionable insights, including learners' individual characteristics (e.g., interests, prior knowledge, skills) and learner-generated game data (e.g., time spent, goals or tasks completed). However, these analyses are usually quite simple: we conducted a review on literature to

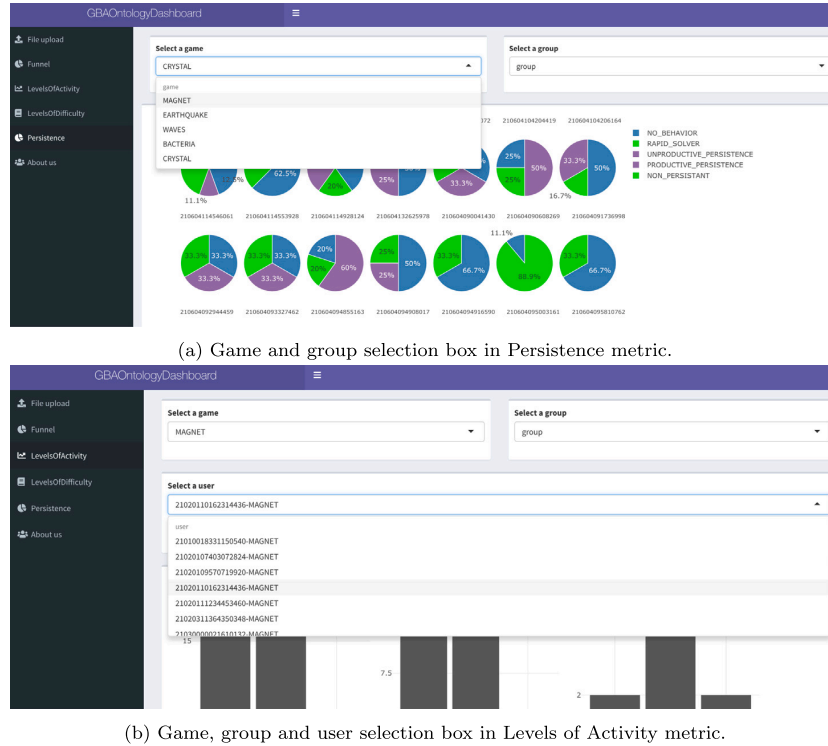


Fig. 6. Selection options in the dashboard.

collect which metrics have been used in previous studies, concluding that most of them only use basic ones (e.g., completion times, count of events, general scores), which implies simple calculations, such as additions or averages. While all these metrics previously developed have been implemented in our research, we have also developed novel metrics using more complex calculations, such as normalization, standardization, and ML algorithms, demonstrating that our ontology can replicate previous literature as well as using new approaches to perform interoperable GBAs.

To see how our environment works in a real context, we collected data from 10 different SGs and used the ontology and metrics to perform GBA and conduct a case study. In many fields, dashboards are used as a tool to inform and transmit knowledge, and their importance and usefulness make them the subject of many studies (Ruipérez-Valiente et al., 2021). Although some studies in the field have used visualizations and deployed dashboards as effective tools to represent GBA data (Gomez, Ruipérez-Valiente, Martínez, & Kim, 2020; Kim, Lin, & Ruipérez-Valiente, 2021), our dashboard designed for the case study is the first one in the area supporting interoperable GBAs, since it covers data from different SGs at the same time by simply using the ontology developed.

This work also has also some limitations: first, data has to be incorporated into the ontology providing the data file manually, and although we have defined a data format with columns that almost any log data from the area should have, probably some adaptations to the original data would be necessary in order to meet the input's requirements. Moreover, this approach supports ML techniques, but it does not support more complex methods, such as Knowledge Inference (KI) or Deep Learning (DL). The use of these methods could help to infer more useful information from learners' data, as well as improving the results' validity and reliability.

We can see the great potential that GBA have applied in many contexts. Regarding professional environments, companies have begun

to use GBA for employee recruitment and selection (Bina, Mullins, & Petter, 2021). In healthcare, games are also being used for assessment, training and rehabilitation (Ferreira-Brito et al., 2019). Moreover, GBA can also be used to measure psychological well-being, by analyzing learners' anxiety, for example (Smits & Charlier, 2011). Despite all these opportunities, the current challenge is (still) to make use of data from learners, teachers, and game learning environments for assessments. Therefore, we firmly believe that the future of games for assessment is promising, and our work can help to alleviate some of the challenges in the area by providing a common knowledge model and provide straightforward interoperable assessments.

6. Conclusions

This research aimed to develop a novel approach to achieve interoperable GBAs using ontologies and in-game metrics that are automatically computed using ontology-based data. With that purpose in mind, we established three objectives: (1) to design and develop an ontology to standardize the GBA area, (2) to conduct a validation study on previous metrics in literature, as well as to design and implement novel metrics, and (3) to conduct a case study illustrating how our approach can be used in a real environment. First, we designed an ontology from scratch using Methontology (Fernández-López et al., 1997) as a base in our methodology, and we conducted a formal evaluation using OOPS! (Poveda-Villalón et al., 2014) to detect possible issues and iterate over the methodology to solve them if necessary. Then, we conducted an study on previous GBA literature, carefully reviewing each paper and noting the metrics developed so that we could replicate them later in our environment. In addition to those metrics, we also designed novel interoperable metrics (including more complex calculations and ML algorithms) to demonstrate and validate the capabilities of our work. Finally, we conducted a case study performing GBA with data from 10 different SGs. Benefiting from the capacities of the framework

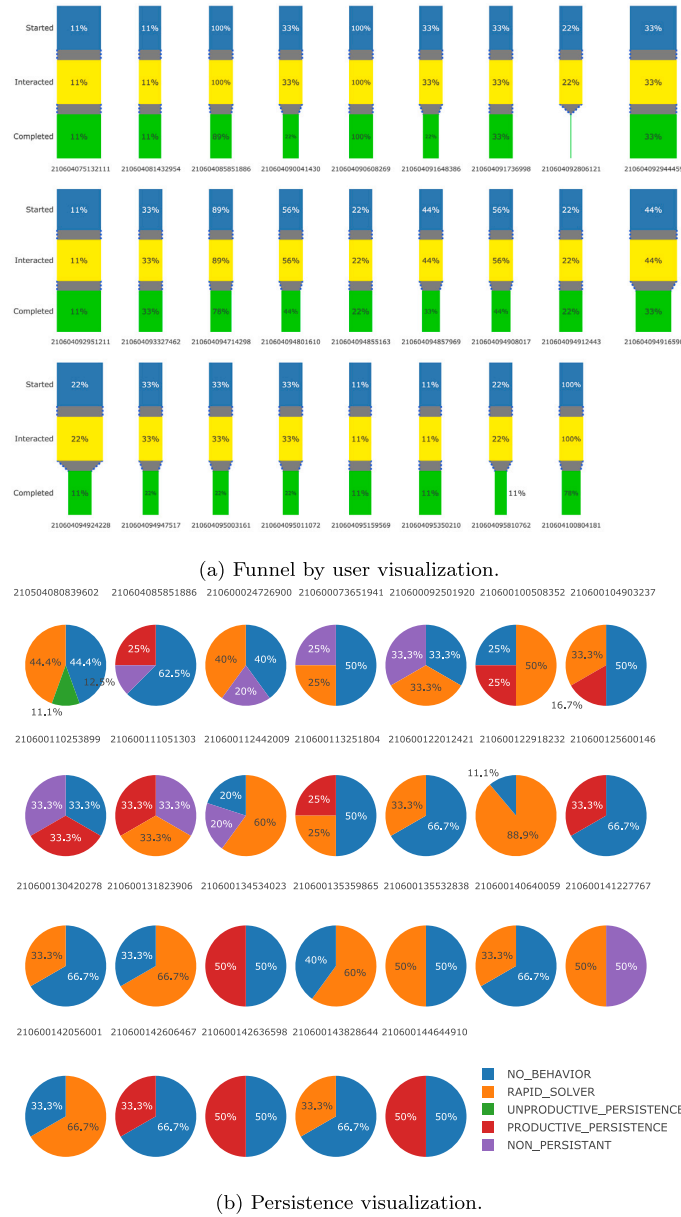
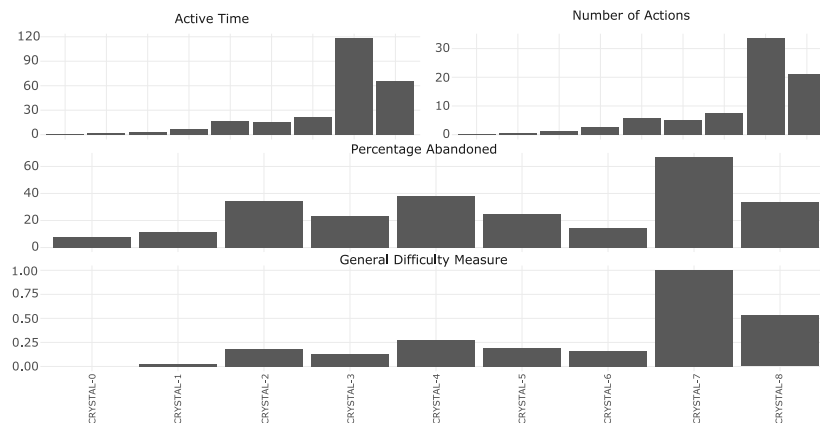


Fig. 7. Funnel by user and Persistence visualizations for the selected game and group.

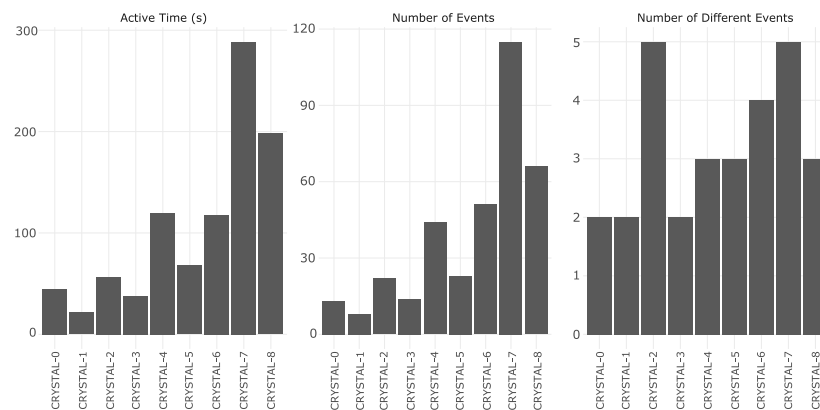
used to integrate our ontology (Gomez et al., 2023), a dashboard was developed using Shiny's R framework and deployed via ShinyApps server to test our approach with real data.

As part of our future work, we would like to validate our approach by conducting case studies (using the framework showed in this research) and collecting data in real time from learners and instructors. Moreover, we will be developing new metrics to continue expanding the system and its possibilities. Future studies should also consider integrating more complex algorithms (KI,DL) that are quite important within the GBA machinery literature in form of metrics that could unveil the full potential that interaction data from learners have in assessment. Finally, novel uses of our ontology and metrics, such as reports generation, will be explored in order to perform new case

studies, giving our research even more practical application. This work provides significant contributions to the literature, including a new ontology designed as a common knowledge model to unify the GBA area, aiming to provide interoperable GBAs using in-game metrics to monitor learners' interaction with games and provide useful insights. In addition, since we developed the metrics in a modular way, integrating them into a powerful ontology-based framework, our work can be easily expanded with new metrics using novel approaches, such as ML techniques. We expect the use of our work (including the ontology and GBA metrics) along with its integration in the framework to solve the current limitations regarding GBA interoperability, reducing the cost and effort of developing specific GBAs, and therefore allowing the deployment of GBAAaaS.



(a) Levels of difficulty visualization for the selected game and group.



(b) Levels of activity visualization for the selected game, group, and user.

Fig. 8. Levels of difficulty and Levels of activity visualizations.

CRedit authorship contribution statement

Manuel J. Gomez: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Visualization. **José A. Ruipérez-Valiente:** Conceptualization, Writing – review & editing, Supervision, Project administration. **Félix J. García Clemente:** Conceptualization, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by (a) grant 21795/FPI/22 - Séneca Foundation. Cofinanced by Innovatiio Global Educación. Region of Murcia (Spain), (b) REASSESS project (grant 21948/JLI/22), funded by the Call for Projects to Generate New Scientific Leadership, included in the Regional Program for the Promotion of Scientific and Technical

Excellence Research (2022 Action Plan) of the Seneca Foundation, Science and Technology Agency of the Region of Murcia, and (c) the strategic project CDL-TALENTUM from the Spanish National Institute of Cybersecurity (INCIBE) and by the Recovery, Transformation and Resilience Plan, Next Generation EU. In addition, we would like to express our gratitude to Field Day Lab for sharing their datasets publicly, which have been used in our experiments.

References

- Abbes, H., & Gargouri, F. (2016). Big data integration: A MongoDB database and modular ontologies based approach. *Procedia Computer Science*, 96, 446–455.
- Abt, C. C. (1987). *Serious games*. University press of America.
- All, A., Castellar, E. P. N., & Van Looy, J. (2014). Measuring effectiveness in digital game-based learning: A methodological review. *International Journal of Serious Games*, 1(2).
- Alonso-Fernandez, C., Calvo, A., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2017). Systematizing game learning analytics for serious games. In *2017 IEEE global engineering education conference* (pp. 1111–1118). IEEE.
- Alonso-Fernandez, C., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, 141, Article 103612.
- Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, 99, 301–309.
- Basu, S., Disalvo, B., Rutstein, D., Xu, Y., Roschelle, J., & Holbert, N. (2020). The role of evidence centered design and participatory design in a playful assessment for computational thinking about data. In *Annual conference on innovation and technology in computer science education* (pp. 985–991). <http://dx.doi.org/10.1145/3328778.3366881>.

- Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: an overview. *Advances in Human-Computer Interaction*, 2013.
- Bina, S., Mullins, J., & Petter, S. (2021). Examining game-based approaches in human resources recruitment and selection: A literature review and research agenda. In *Proceedings of the 54th hawaii international conference on system sciences* (p. 1325).
- Bürger, T., & Simperl, E. (2008). Measuring the benefits of ontologies. In *On the move to meaningful internet systems: OTM 2008 workshops: oTM confederated international workshops and posters, ADI, aWeSoMe, COMBEK, EI2N, IWSSA, MONET, onToContent+ QSI, ORM, perSys, RDDS, SEMELS, and SWWS 2008, Monterrey, Mexico, November 9-14, 2008. proceedings* (pp. 584–594). Springer.
- Card, M. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.
- Chiu, F.-Y., & Hsieh, M.-L. (2017). Role-playing game based assessment to fractional concept in second grade mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(4), 1075–1083. <http://dx.doi.org/10.12973/eurasia.2017.00659a>.
- Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, 46(1), 41–64.
- Daniel, M., & Garry, C. (2018). *Video games as culture: considering the role and importance of video games in contemporary society*. Routledge.
- De Freitas, S. (2006). Learning in immersive worlds: A review of game-based learning. *JISC ELearning Innov.*
- de Klerk, S., & Kato, P. M. (2017). The future value of serious games for assessment: Where do we go now? *Journal of Applied Testing Technology*, 18(S1), 32–37.
- Dehainsala, H., Pierra, G., & Bellatreche, L. (2007). Ontodb: An ontology-based database for data intensive applications. In *International conference on database systems for advanced applications* (pp. 497–508). Springer.
- Díaz, A. R., Benito-Santos, A., Dorn, A., Abgaz, Y., Wandl-Vogt, E., & Therón, R. (2019). Intuitive ontology-based SPARQL queries for RDF data exploration. *IEEE Access*, 7, 156272–156286.
- DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society*, 17(1), 17–28.
- Donovan, L., & Lead, P. (2012). The use of serious games in the corporate sector. In *A state of the art report*. Learnovate Centre (December 2012).
- Dou, D., Wang, H., & Liu, H. (2015). Semantic data mining: A survey of ontology-based approaches. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing* (pp. 244–251). IEEE.
- Duval, E. (2011). Attention please! learning analytics for visualization and recommendation. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 9–17).
- ESA (2021). *2021 Essential facts about the computer and video game industry: Technical report*, Entertainment Software Association.
- Fathy, N., Gad, W., & Badr, N. (2019). A unified access to heterogeneous big data through ontology-based semantic integration. In *2019 ninth international conference on intelligent computing and information systems* (pp. 387–392). IEEE.
- Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 spring symposium* (pp. 33–40).
- Ferreira-Brito, F., Fialho, M., Virgolino, A., Neves, I., Miranda, A. C., Sousa-Santos, N., et al. (2019). Game-based interventions for neuropsychological assessment, training and rehabilitation: Which game-elements to use? A systematic review. *Journal of Biomedical Informatics*, 98, Article 103287.
- Field Day Lab (2022). We're field day. URL <https://fielddaylab.wisc.edu/about/>. (Last Accessed on 29 April 2022).
- Gagnon, D. J., & Swanson, L. (2023). Open game data: A technical infrastructure for open science with educational games. In *Joint international conference on serious games* (pp. 3–19). Springer.
- Gandon, F., Bottollier, V., Corby, O., & Durville, P. (2007). RDF/XML source declaration.
- Georgiou, K., Gouras, A., & Nikolaou, I. (2019). Gamification in employee selection: The development of a gamified assessment. *International Journal of Selection and Assessment*, 27(2), 91–103.
- Girard, C., Ecalte, J., & Magnan, A. (2013). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, 29(3), 207–219.
- Gomez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies*.
- Gomez, M. J., Ruipérez-Valiente, J. A., & García Clemente, F. J. (2023). A framework to support interoperable game-based assessments as a service (GBAaaS): Design, development, and use cases. *Software - Practice and Experience*, 53(11), 2222–2240.
- Gomez, M. J., Ruipérez-Valiente, J. A., Martínez, P. A., & Kim, Y. J. (2020). Exploring the affordances of sequence mining in educational games. In *Eighth international conference on technological ecosystems for enhancing multicultural* (pp. 648–654).
- Gris, G., & Bengtson, C. (2021). Assessment measures in game-based learning research: a systematic review. *International Journal of Serious Games*, 8(1), 3–26.
- Gros, B. (2007). Digital games in education: The design of games-based learning environments. *Journal of Research on Technology in Education*, 40(1), 23–38.
- Hamdaoui, N., Khalidi Idrissi, M., & Bennani, S. (2016). Adaptive educational games using game metrics. In *International afro-European conference for industrial advancement* (pp. 198–208). Springer.
- Happel, H.-J., & Seedorf, S. (2006). Applications of ontologies in software engineering. In *Proc. of workshop on semantic web enabled software engineering (SWESE) on the ISWC* (pp. 5–9). Citeseer.
- Hinton, S. (2016). Applying gamification in New Zealand contact centers. *Special Interest Group on Human-Computer Interaction*.
- Ifenthaler, D., Eseryel, D., & Ge, X. (2012). Assessment for game-based learning. In *Assessment in game-based learning* (pp. 1–8). Springer.
- Jayapandian, C. P., Zhao, M., Ewing, R. M., Zhang, G.-Q., & Sahoo, S. S. (2012). A semantic proteomics dashboard (SemPoD) for data management in translational research. *BMC Systems Biology*, 6(3), 1–13.
- Kangas, M., Koskinen, A., & Krokfors, L. (2017). A qualitative literature review of educational games in the classroom: the teacher's pedagogical activities. *Teachers and Teaching*, 23(4), 451–470.
- Kato, P. M., & de Klerk, S. (2017). Serious games for assessment: Welcome to the jungle. *Journal of Applied Testing Technology*, 18(S1), 1–6.
- Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In *Game-based assessment revisited* (pp. 3–11). Springer.
- Kim, Y. J., Lin, G., & Ruipérez-Valiente, J. A. (2021). Expanding teacher assessment literacy with the use of data visualizations in game-based assessment. In *Visualizations and dashboards for learning analytics* (pp. 399–419). Springer.
- Laamarti, F., Eid, M., & El Saddik, A. (2014). An overview of serious games. *International Journal of Computer Games Technology*, 2014.
- Larson, K. (2020). Serious games and gamification in the corporate training environment: A literature review. *TechTrends*, 64(2), 319–328.
- Lee, J., Luchini, K., Michael, B., Norris, C., & Soloway, E. (2004). More than just fun and games: Assessing the value of educational video games in the classroom. In *CHI'04 extended abstracts on human factors in computing systems* (pp. 1375–1378).
- Lehmann, J., Sejdíu, G., Bühlmann, L., Westphal, P., Stadler, C., Ermilov, I., et al. (2017). Distributed semantic analytics using the SANS stack. In *International semantic web conference* (pp. 147–155). Springer.
- Liu, M., Kang, J., Liu, S., Zou, W., & Hodson, J. (2017). Learning analytics as an assessment tool in serious games: A review of literature. *Serious Games and Edutainment Applications*, 537–563.
- Loh, C. S., & Sheng, Y. (2015). Measuring expert performance for serious games analytics: From data to insights. In *Serious games analytics* (pp. 101–134). Springer.
- Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: vol. 10*, (pp. 3–29). Cham: Springer International Publishing.
- Martínez, P. A., Gómez, M. J., Ruipérez-Valiente, J. A., Pérez, G. M., & Kim, Y. J. (2020). Visualizing educational game data: A case study of visualizations to support teachers.
- McDaniel, M., & Storey, V. C. (2019). Evaluating domain ontologies: clarification, classification, and challenges. *ACM Computing Surveys*, 52(4), 1–44.
- Mollick, E., & Werbach, K. (2015). Gamification and the enterprise. *The Gameful World: Approaches, Issues, Applications*, 439.
- Nikolaou, I., Georgiou, K., & Kotsasarlidou, V. (2019). Exploring the relationship of a gamified assessment with performance. *Spanish Journal of Psychology*, <http://dx.doi.org/10.1017/sjp.2019.5>.
- Nouira, A., Cheniti-Belcadhi, L., & Braham, R. (2018). An enhanced xAPI data model supporting assessment analytics. *Procedia Computer Science*, 126, 566–575.
- Olzewska, J. I., Houghtaling, M., Goncalves, P. J., Fabiano, N., Haidegger, T., Carbonera, J. L., et al. (2020). Robotic standard development life cycle in action. *Journal of Intelligent and Robotic Systems*, 98(1), 119–131.
- Owen, V. E., & Baker, R. S. (2019). Learning analytics for games. *Handbook of game-based learning*, 513–535.
- Panov, P., Džeroski, S., & Soldatova, L. (2008). OntoDM: An ontology of data mining. In *2008 IEEE international conference on data mining workshops* (pp. 752–760). IEEE.
- Perez-Colado, V. M., Rotaru, D. C., Freire, M., Martínez-Ortiz, I., & Fernandez-Manjon, B. (2018). Learning analytics for location-based serious games. In *2018 IEEE global engineering education conference* (pp. 1192–1200). IEEE.
- Petri, G., & von Wangenheim, C. G. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education*, 107, 68–90.
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283.
- Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., et al. (2013). Metrics in simulations and games for learning. In *Game analytics* (pp. 697–729). Springer.
- Podgorelec, V., & Kuhar, S. (2011). Taking advantage of education data: Advanced data analysis and reporting in virtual learning environments. *Elektronika ir Elektrotehnika*, 114(8), 111–116.
- Poveda-Villalón, M., Gómez-Pérez, A., & Suárez-Figueroa, M. C. (2014). Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2), 7–34.
- al Qallawi, S., & Raghavan, M. (2022). A review of online reactions to game-based assessment mobile applications. *International Journal of Selection and Assessment*, 30(1), 14–26.

- Qian, M., & Clark, K. R. (2016). Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, 63, 50–58.
- Reyes-Chua, E., & Lidawan, M. W. (2019). Games as effective language classroom strategies: a perspective from english major students. *European Journal of Foreign Language Teaching*.
- Rocha, O. R., & Zucker, C. F. (2015). Ludo: an ontology to create linked data driven serious games. In *ISWC 2015-workshop on LINKed eDucation, LINKED 2015, at Bethlehem, Pennsylvania, United states*.
- Ruiperez-Valiente, J. A., Gaydos, M., Rosenheck, L., Kim, Y. J., & Klopfer, E. (2020). Patterns of engagement in an educational massively multiplayer online game: A multidimensional view. *IEEE Transactions on Learning Technologies*, 13(4), 648–661.
- Ruipérez-Valiente, J. A., Gomez, M. J., Martínez, P. A., & Kim, Y. J. (2021). Ideating and developing a visualization dashboard to support teachers using educational games in the classroom. *IEEE Access*, 9, 83467–83481.
- Said, B., Cheniti-Belcadhi, L., & El Khayat, G. (2019). An ontology for personalization in serious games for assessment. In *2019 IEEE second international conference on artificial intelligence and knowledge engineering* (pp. 148–154). IEEE.
- Santos, H., Dantas, V., Furtado, V., Pinheiro, P., & McGuinness, D. L. (2017). From data to city indicators: A knowledge graph for supporting automatic generation of dashboards. In *European semantic web conference* (pp. 94–108). Springer.
- Serrano, Á., Marchiori, E. J., del Blanco, Á., Torrente, J., & Fernández-Manjón, B. (2012). A framework to improve evaluation in educational games. In *Proceedings of the 2012 IEEE global engineering education conference* (pp. 1–8). IEEE.
- Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces*, 50, 116–123.
- Shoukry, L. (2020). *Mobile multimodal serious games analytics* (Ph.D. thesis), Technische Universität.
- Shoukry, L., Göbel, S., & Steinmetz, R. (2014). Learning analytics and serious games: Trends and considerations. In *Proceedings of the 2014 ACM international workshop on serious games* (pp. 21–26).
- Shute, V. J., & Ventura, M. (2015). Stealth assessment. *The SAGE Encyclopedia of Educational Technology*, 675–676.
- Sliney, A., & Murphy, D. (2011). Using serious games for assessment. In *Serious games and edutainment applications* (pp. 225–243). Springer.
- Smits, J., & Charlier, N. (2011). Game-based assessment and the effect on test anxiety: A case study. In *European conference on games based learning* (p. 562). Academic Conferences International Limited.
- Song, Y., & Sparks, J. (2019). Measuring argumentation skills through a game-enhanced scenario-based assessment. *Journal of Educational Computing Research*, 56(8), 1324–1344. <http://dx.doi.org/10.1177/0735633117740605>.
- Staab, S., & Studer, R. (2010). *Handbook on ontologies*. Springer Science & Business Media.
- Stănescu, I. A., Stefan, A., Kravcik, M., Lim, T., & Bidarra, R. (2013). Interoperability strategies for serious games development. *Internet Learning*, 2(1), 6.
- Sun, Z., Hu, C., Li, C., & Wu, L. (2020). Domain ontology construction and evaluation for the entire process of software testing. *IEEE Access*, 8, 205374–205385.
- Susi, T., Johannesson, M., & Backlund, P. (2007). *Serious games: An overview: Technical report*, Institutionen för kommunikation och information.
- Tang, S., & Hanneghan, M. (2011). Game content model: an ontology for documenting serious game design. In *2011 developments in e-systems engineering* (pp. 431–436). IEEE.
- Theodosiou, S., & Karasavvidis, I. (2015). Serious games design: A mapping of the problems novice game designers experience in designing games. *Journal of e-Learning and Knowledge Society*, 11(3).
- Tudorache, T., Vendetti, J., & Noy, N. F. (2008). Web-protege: A lightweight OWL ontology editor for the web. In *OWLED: vol. 432*, (p. 2009).
- Uschold, M. (1996). Building ontologies: Towards a unified methodology. In *Proceedings of 16th annual conference of the british computer society specialists group on expert systems*. Citeseer.
- Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, 11(2), 93–136.
- Valiente, J. A. R. (2022). Unveiling the potential of learning analytics in game-based learning: Case studies with a geometry game. In *Handbook of research on promoting economic and social development through serious games* (pp. 524–544). IGI Global.
- Vallejo, V., Wyss, P., Rampa, L., Mitache, A. V., Müri, R. M., Mosimann, U. P., et al. (2017). Evaluation of a novel Serious Game based assessment tool for patients with Alzheimer's disease. *PLoS One*, 12(5), Article e0175999.
- Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10), 1500–1509.
- Wiemeyer, J., & Kliem, A. (2012). Serious games in prevention and rehabilitation—a new panacea for elderly people? *European Review of Aging and Physical Activity*, 9(1), 41–50.
- Yusoff, A., Crowder, R., Gilbert, L., & Wills, G. (2009). A conceptual framework for serious games. In *2009 ninth IEEE international conference on advanced learning technologies* (pp. 21–23). IEEE.

3 A Framework for Interoperable GBaaS

Title

A framework to support interoperable Game-based Assessments as a Service (GBaaS): Design, development, and use cases

Authors

Manuel J. Gomez¹, José A. Ruipérez-Valiente¹,
Félix J. García Clemente¹

¹*Department of Information and Communications Engineering,
University of Murcia, Spain*

Publication details

| | | | |
|----------------|-----------------------------------|------------------|------------------|
| Journal | Software: Practice and Experience | Publisher | Wiley |
| Volume | 53 | Number | 11 |
| Pages | 2222-2240 | Year | 2023 |
| JIF | 2.6 | Rank | Q2 |
| Status | Published | DOI | 10.1002/spe.3254 |

Abstract

During the last few years, there has been increasing attention paid to serious games (SGs), which are games used for non-entertainment purposes. SGs offer the potential for more valid and reliable assessments compared to traditional methods such as paper-and-pencil tests. However, the incorporation of assessment features into SGs is still in its early stages, requiring specific design efforts for each game and adding significant time to the design of Game-based Assessments (GBAs). In this research, we present a completely novel framework that aims to perform interoperable GBAs by: (a) integrating a common GBA ontology model to process RDF data; (b) developing in-game metrics to infer useful information and assess learners; (c) integrating a service API to provide an easy way to interact with the framework. We then validate our approach through performance evaluation and two use cases, demonstrating its effectiveness in real-world scenarios with large-scale datasets. Our results show that the developed framework achieves excellent performance, replicating metrics from previous literature. We anticipate that our work will help alleviate current limitations in the field and facilitate the deployment of GBAs as a Service.



A framework to support interoperable Game-based Assessments as a Service (GBAaaS): Design, development, and use cases

Manuel J. Gomez¹ | José A. Ruipérez-Valiente² | Félix J. García Clemente³

Faculty of Computer Science, University of Murcia, Murcia, Spain

Correspondence

Manuel J. Gomez, Faculty of Computer Science, University of Murcia, Calle Campus Universitario 32, 30100, Murcia, Spain.
Email: manueljesus.gomez@um.es

Funding information

Fundación Séneca, Grant/Award Number: 21795/FPI/22; Call for Projects to Generate New Scientific Leadership, Grant/Award Number: 21948/JLI/22

Abstract

During the last few years, there has been increasing attention paid to serious games (SGs), which are games used for non-entertainment purposes. SGs offer the potential for more valid and reliable assessments compared to traditional methods such as paper-and-pencil tests. However, the incorporation of assessment features into SGs is still in its early stages, requiring specific design efforts for each game and adding significant time to the design of Game-based Assessments (GBAs). In this research, we present a completely novel framework that aims to perform interoperable GBAs by: (a) integrating a common GBA ontology model to process RDF data; (b) developing in-game metrics to infer useful information and assess learners; (c) integrating a service API to provide an easy way to interact with the framework. We then validate our approach through performance evaluation and two use cases, demonstrating its effectiveness in real-world scenarios with large-scale datasets. Our results show that the developed framework achieves excellent performance, replicating metrics from previous literature. We anticipate that our work will help alleviate current limitations in the field and facilitate the deployment of GBAs as a Service.

KEYWORDS

big data, data mining, educational technology, Game-based Assessment, interoperability, ontologies

1 | INTRODUCTION

Video games have assumed an essential place in our lives, evolving into complex and diverse platforms that are enjoyed by people of all ages and backgrounds.^{1,2} This has generated increasing interest in using games in various settings during the last decade.³ The application of games with a non-entertainment primary purpose, known as serious games (SGs), can provide multiple benefits in environments where games were not traditionally used.⁴ Education is one such domain,

Abbreviations: AI, artificial intelligence; API, application programming interface; GBAs, game-based assessments; GBAaaS, game-based assessment as a service; HDFS, hadoop distributed file system; ML, machine learning; OBDA, ontology-based data access; OWL, web ontology language; RDF, resource description framework; RDD, resilient distributed dataset; REST, representational state transfer; SANSa, scalable semantic analytics stack; SGs, serious games.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Software: Practice and Experience* published by John Wiley & Sons Ltd.

where traditional content still constitute the majority of learning materials, and there is little consensus on how to effectively integrate technology in the classrooms.⁵ However, SGs are also used in many other domains,⁶ including training, well-being, advertisement, and interpersonal communication, among others. SGs are also being explored as assessment tools, in particular for their potential to provide more valid assessments compared to traditional assessment approaches, providing more meaningful and authentic contexts for assessments through interactive and immersive environments.⁷ In the context of Game-based Assessment (GBA), a key challenge is making valid inferences about what the student knows, believes, and can do without disrupting the game flow.⁸

Furthermore, GBA machinery, which includes metrics, dashboards, and other analytics,^{9,10} is usually designed for each game, leading to increased costs, time, and effort. This development also requires the maintenance of a complete infrastructure, requiring dedicated engineers to address these tasks. Semantic web technologies, and particularly ontologies, can address these heterogeneity problems. Ontologies capture domain-specific knowledge and offer an explicit common conceptualization.¹¹ Nowadays, ontologies are used in various application areas involving artificial intelligence, natural language processing, data integration, and knowledge management.¹² Using ontologies to define a standard knowledge model in the context of Game-based Assessments (GBAs) could alleviate the time-consuming and costly step of creating GBAs.¹³

The growing use of GBAs has led to the creation of large data repositories, presenting new assessment opportunities.^{14,15} A game (educational or not) can generate vast amounts of interaction data, even in a short game-play session. Data mining and visualization techniques applied to player interaction logs can provide valuable insights into how players engage with the game, leading to improvements in assessment methods and real-time feedback on activity progress.¹⁶ However, data processing usually shows performance deficiencies when the dataset exceeds the memory size of a single machine, and distributed computing frameworks can be employed to address this limitation.^{17,18} Developing an efficient system capable of processing large amounts of data and performing interoperable GBAs could significantly simplify the design process. In addition, utilizing this system as an external service could reduce costs and efforts, enabling the use of GBAs as a Service (GBAaaS).

In this research, we present a novel approach that combines the use of ontologies with big data technologies to create interoperable GBAs. To address the challenge of game interoperability and specific assessment machinery, we have developed a new framework that automatically computes in-game metrics using the provided data and the standard ontology model, where we re-use the existing Scalable Semantic Analytics Stack (SANSA).¹⁹ Additionally, to solve the infrastructure development and maintenance issues, we integrate a service API into our framework, allowing it to function as an external service and enabling the GBAaaS paradigm. Specifically, in this research, we present the following contributions:

- **Framework development.** Using SANSA-Stack as a baseline, we develop a framework that integrates our created GBA ontology using big data technologies. This facilitates game interoperability and enables the GBAaaS paradigm. The main novel features of our framework are:
 - **Ontology integration.** Our framework integrates the previously designed ontology into the SANSA framework to process large resource description framework (RDF) data.
 - **In-game metric development.** We develop and integrate a set of basic GBA metrics from the literature into the framework, enabling interoperable learners' assessment using RDF data. These metrics are also used to test the framework's performance when querying large datasets.
 - **Service API integration.** We develop and integrate a REST API into our ecosystem, providing an easy way to interact with the framework and allowing the use of our framework as an input/output service (GBAaaS).
- **Framework evaluation.** We have conducted a performance evaluation and a case study validation:
 - **Performance evaluation.** This involves two tasks aiming to validate our framework's input services: (1) data scalability, testing how our framework scales to larger datasets and what the improvement is concerning the number of workers in the cluster mode; and (2) flexibility, testing how the framework processes different metrics.
 - **Case study validation.** We present a case study with two use cases to demonstrate how our framework can be used as GBAaaS in various real applications of GBA in educational environments.

The rest of the article is structured as follows: Section 2 reviews background literature on SGs, GBA, and the use of ontologies in big data environments. Section 3 presents the framework proposal and the case studies performed to test

its performance and capabilities. Then, we finalize the article with a discussion in Section 5, and conclusions and future work in Section 6.

2 | RELATED WORK

In this section, we present a review of the literature in the areas most closely related to our work: in Section 2.1, we review literature related to SGs and GBA, and Section 2.2 reviews the literature related to ontologies and their use in big data environments.

2.1 | Serious games and game-based assessment

The idea of playing a game dates to the ancient past and is considered an integral part of all societies.⁶ In addition to the previously mentioned benefits, it is argued that SGs can also positively impact the players' development of several different skills.²⁰ SGs are currently being used in several contexts. For example, using games for formal education has become widely accepted as playing games has become an essential part of young people's lives worldwide.²¹ Additionally, there is also an increasing interest in how games can be effectively applied in learning and training contexts, as well as in other areas such as healthcare,²² rehabilitation,²³ and military training.²⁴ For instance, Albaladejo et al.²⁵ presented a multimodal system that could be used to improve cyberdefense capabilities by utilizing gamified platforms for training and testing individuals and organizations in cybersecurity practices and techniques.

However, SGs must be able to show that the necessary learning has occurred. An advantage of SGs as assessment tools is that they can be programmed to capture, store, and share massive amounts of user data over time.⁷ This data can be used to perform reliable assessments and manifest this learning, enabling GBA. A common approach to perform this assessment is to use a set of metrics (or indicators) that transform raw data into meaningful information. For example, the authors in Reference 26 proposed a multidimensional measurement of engagement in a learning game ("The Radix Endeavor") across four dimensions: general activity, social, exploration, and quests. Similarly, researchers in Reference 27 explored the creation of engagement profiles based on log data, considering the different ways players engage with the game and highlighting patterns associated with active play. Many other studies have conducted research aiming to assess users' interaction with games, measuring factors such as persistence,²⁸⁻³⁰ difficulty,^{31,32} and level completion,^{10,33} among other measures. Furthermore, we found examples of metrics developed in non-educational contexts. Authors in Reference 34 developed a GBA using multi-level functional tasks to assess instrumental activities of daily living in a sample of inpatients with chronic schizophrenia, measuring completion times and errors in each task. Furthermore, Jackson et al.³⁵ assessed a sample of 67 Reserve Officers' Training Corps (military sample) using the commercial game "Crysis 2,"³⁶ which simulates key features involved in combat situations. In this research, authors measured tasks such as the number of eliminations, shots accuracy, or damage per bullet, among others.

As we have seen, many SGs track their learners' interactions, but they usually use custom formats.³⁷ However, there are enough case studies that identify common interactions tracked by SGs to start defining a standardized model. Previous studies have proposed approaches to standardize analytics in games. For example, Serrano et al.³⁷ presented xAPI, an implementation of a standard model that sets a basis for performing analysis in SGs methodologically. Moreover, Said et al.³⁸ proposed an ontology to model player experience and its association with in-game personalization, also defining a set of reasoning rules to suggest tailored games for each player's assessment path and player experience. Authors in Reference 39 proposed an ontology that allows the description and representation of SGs that use resources from the Web of Data, introducing concepts such as "game structure," "game simulation," or "game rule." However, we did not find any study attempting to standardize the GBA area to build interoperable assessments.

In this research, we go beyond the existing literature by attempting to standardize the GBA area through the definition of a common knowledge model. With this purpose in mind, we combine ontologies, big data technologies, metrics, and API services to create a novel framework capable of analyzing and inferring new knowledge using data from different games. This information is analyzed and transformed using interoperable metrics, which are available for consultation in various output formats. Finally, the API service has been defined and integrated to facilitate interaction with our framework by different sources.

2.2 | Ontologies and big data architectures

Over the last 15 years, ontology-based applications have been spreading and maturing. One can now find ontology-based applications in diverse areas, including customer support and car engineering.⁴⁰ For instance, researchers in Reference 41 designed and developed an integrated ontology of software engineering approaches to support sustainable software development knowledge, awareness, and implementation. Furthermore, in Reference 42, authors conducted an analysis that explained a detailed approach to building an ontology that can be used across different e-learning platforms. Additionally, as we have seen in the previous section, several studies^{37-39,43} presented ontologies in different SGs applications (e.g., collaborative learning, Web Data technologies). However, little research has been conducted in the direction on ontology development for GBA. To the best of our knowledge, no ontology has been proposed specifically for the GBA area, except for the one we presented in our previous research.⁴⁴ This ontology is part of the framework and acts as the intermediate semantic layer in this research.

Several tools for managing ontology and ontology-based data are available, such as Protégé. Previous research includes studies and frameworks that dealt with ontology-based data. Botoeva et al.⁴⁵ generalized ontology-based data access (OBDA) to allow querying arbitrary data sources using SparQL and compared implementing an OBDA system over MongoDB with a triple store. Moreover, authors in Reference 46 presented *Minerva*, a storage and inference system for large-scale OWL ontologies on top of relational databases. *Minerva* comprises four different modules: an import module for reading ontology data, an inference module, a storage module for storing original and inferred assertions, and a query module that uses SparQL.

However, storing ontology data in the computer's main memory is a problem for applications that manipulate a large amount of ontology-based data.⁴⁷ In recent years, several studies have proposed new approaches that use big data technologies to manage ontology-based data. For example, Abbes and Gargouri⁴⁸ proposed an approach based on MongoDB and modular ontologies. They made it possible by wrapping data sources to MongoDB databases, generating local ontologies, and finally composing the local ontologies to get a global one. Mountasser et al.⁴⁹ presented a semantic-based big data integration framework that relies on large-scale ontology matching and probabilistic-logical-based assessment strategies; they proved its efficiency in terms of accuracy, performance, and scalability. Moreover, Reyes-Álvarez et al.⁵⁰ presented a novel approach that enables the distributed storage of ontology-based data by exploiting the inherent distribution of NoSQL database nodes. Finally, authors in Reference 19 presented SANSA, a big data engine for scalable processing of large-scale RDF data using Spark and Flink. In our research, we use SANSA as a base for our framework. In particular, we take advantage of various SANSA functionalities, such as the "Read/Write RDF library," and the "inference library." Furthermore, we enhance SANSA's functionalities by adding custom rules and queries to infer new knowledge from existing data, support for various output formats, and a service API.

3 | FRAMEWORK PROPOSAL

3.1 | Framework requirements

Classic assessment has evolved over the past several years from traditional pen and paper-based tests to the use of technology, such as games, and continues to evolve.⁵¹ However, implementing assessment features into games is only in its early stages because it adds a time-consuming step to the design process.¹³ This is due to heterogeneity issues since assessment mechanisms are usually explicitly designed for each game. Moreover, with the challenges brought on by GBAs, including data analytics, the large amount of data now available for teachers is far too complex for conventional database software to store, manage, and process.⁵² Finally, integrating GBAs in different environments generates a diverse range of data from various sources,³⁷ emphasizing the need for a unified and secure way to access GBA data. Therefore, we identified the following requirements:

Requirement 1—Semantic layer between the event data and a common knowledge model: As previous literature reported,^{11,53,54} there are heterogeneity issues with the collected data. In fact, most previous studies did not report any specific format for the collected data. Therefore, we need to define a standard knowledge model that unifies the GBA area and can represent the necessary information for user assessment.

Requirement 2—Processing of large scale data: User interaction with games generate massive amounts of data to be analyzed. Although the authors in Reference 53 reported relatively low sample sizes in the studies, Gomez et al.⁵⁵ stated that GBA research would benefit from using larger datasets since nearly half of the studies in their review described

limitations in data sampling. Thus, it is important to use larger data samples, and we need to be prepared to process large quantities of data to extract useful and reliable information from them.

Requirement 3—Game interoperability for GBA metrics and visualizations: GBA studies previously conducted normally used ad-hoc solutions that enable data gathering to perform some assessment based on users' data. In addition, some studies also enabled visualization dashboards for instructors and users.¹⁰ These solutions had to be developed specifically for each game, severely limiting reusability. Therefore, to scale up the number of GBA implementations, we need to provide new interoperable approaches that can reduce the effort required to build new GBAs, using scalable and interoperable modules that can easily be added and reconfigured.

Requirement 4—Easy communication with external sources: Due to the integration of GBAs in different environments, numerous data sources can hold valuable information to assess user interactions with games. Since many different sources could use this computed information, we need to enable the system to be used as an input/output service to deposit data from different sources or query the interoperable assessments generated. This need could be met by integrating an API that can be used across applications, allowing the use of GBaaS.

Requirement 5—Privacy, authentication, and authorization configurations: Users' privacy in web services is essential to address. GBA data contains valuable but also sensitive information, and sharing or analyzing these data introduces privacy risks for the data subjects, primarily students.⁵⁶ Among the many methods proposed in the literature, role-based access control (RBAC) has been widely accepted as the most promising model because of its simplicity, flexibility in capturing dynamic requirements, and support for the principle of least privilege and efficient privilege management.⁵⁷ Therefore, we need to provide a system with different roles and permissions to ensure that only the appropriate users can access specific analyses and data.

3.2 | Architecture

In this section, we describe each module of the framework's architecture, which can be seen in Figure 1. We divide our framework into five different modules: (1) Preprocessing module, (2) analytics, inference, and querying module, (3) metric output module, (4) authentication and authorization module, and (5) API module. As we can see, the first module aims to transform the raw data (CSV, TSV, JSON ...) into RDF/XML files by using an ontology model. In the next step, we use the SANS framework to process the ontology data, infer new information, and perform queries over the inferred data. In the metric output module, we aim to provide different output formats for the query results. Then, the authentication and authorization module manage the different roles and authorizations in the system, ensuring that clients making requests through our API module are correctly logged in and have the necessary permissions. Finally, the API module aims to facilitate easy integration with external sources by generating a web service that can be used across different applications.

3.2.1 | Preprocessing

The first step is to transform the raw data, which can be received in multiple formats (e.g., CSV, TSV, JSON), into a format understandable by the common knowledge model. In this research, we use the "GBA ontology," a previously developed ontology used as an intermediate knowledge model between the raw data and the metrics outputted by the framework. Since there was no existing ontology that met our requirements and could be expanded, the "GBA ontology" was built from scratch using methontology.⁵⁸ Methontology is a structured method designed to build ontologies from scratch or by re-engineering existing ones.

Our ontology aims to satisfy **R1**, creating an intermediate layer to transform log data (produced by users' interaction with games) into ontology data. This new format is used to analyze and infer new information for assessing users. The ontology includes core concepts required for user assessment given any data set, such as "game event," "attempt," "unit of play" (which is equivalent to a level), "user," or "user group." An overview of the ontology (including classes and relationships between them) can be seen in Figure 2. The reader can view the full ontology using the *Web-Protégé* web page link or the source file, both available in Reference 59. In addition, the full development process and a more detailed view of the ontology can be found in Reference 44.

Using this model, the raw data is transformed into ontology classes, annotations, and relationships that contain the same information. This information is stored using RDF/XML format. The Extensible Markup Language (XML) has become widely adopted, along with transformation languages like XSLT and XQuery, to translate data from one XML format into another.⁶⁰ However, RDF has become another popular data representation and exchange standard. It is a

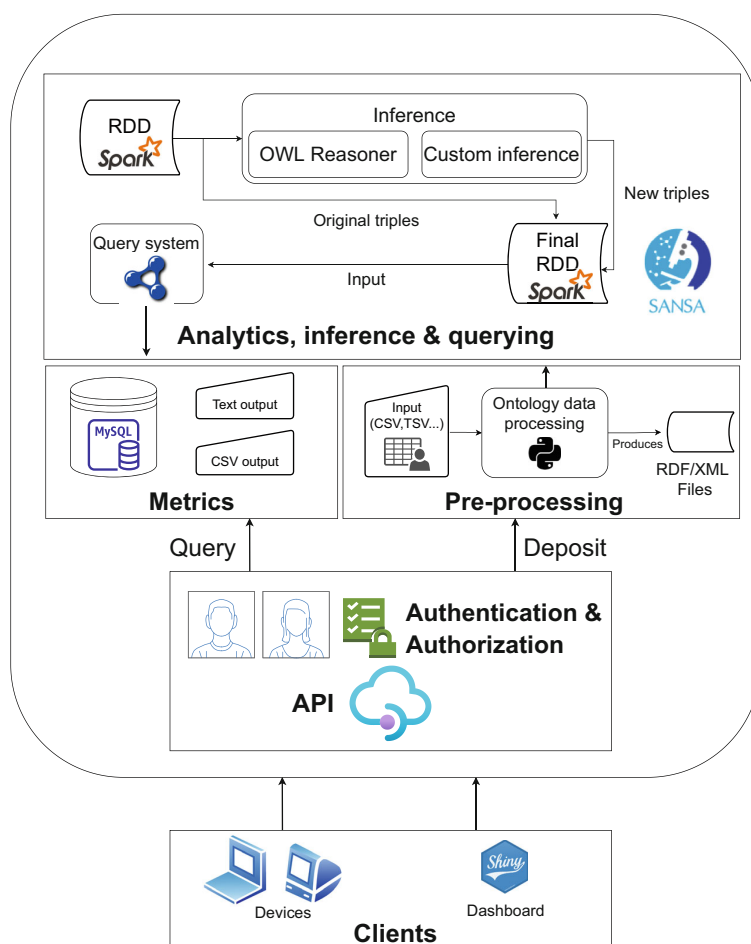


FIGURE 1 Framework's architecture.

general-purpose language for representing data and metadata on the web and is supported by its own query language, SparQL, which enables the extraction and transformation of RDF data. RDF has an XML syntax called RDF/XML, where the formal grammar for the syntax is annotated with actions generating triples of the RDF graph.⁶¹

3.2.2 | Analytics, inference, and querying

Once the data is stored in RDF/XML format, the next step is to process this data and infer new information. To fulfill **R2** and **R3**, we have decided to use the SANSAs framework as the basis for performing our analyses. SANSAs is an open-source structured data processing engine that enables distributed computation over large-scale RDF datasets. It provides data distribution, scalability, and fault tolerance for manipulating large RDF datasets. SANSAs facilitates scalable analytics on the data by utilizing cluster-based big data processing engines, with Spark being the specific engine we employ.¹⁹ An overview of SANSAs architecture is shown in Figure 3. Specifically, SANSAs includes:

- Specialized serialization mechanisms and partitioning schemas for RDF, using vertical partitioning strategies.
- A scalable query engine for large RDF datasets and different distributed representation formats for RDF.
- An adaptive reasoning engine that derives an efficient execution and evaluation plan from a given set of inference rules.
- Several distributed structured machine learning (ML) algorithms can be applied to large-scale RDF data.
- A framework with a unified API that aims to combine distributed in-memory computation technology with semantic technologies.

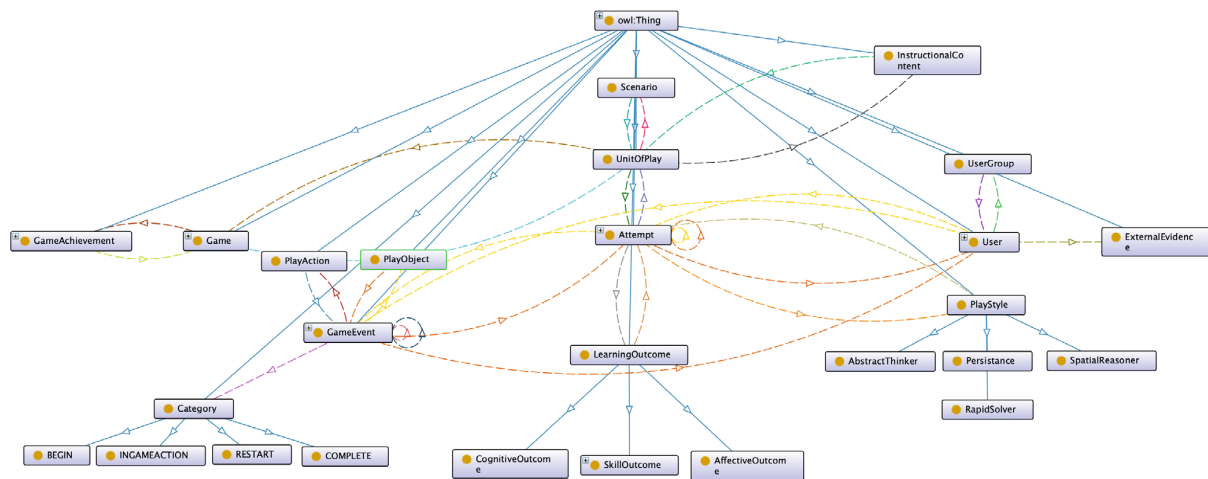


FIGURE 2 An overview of GBA ontology classes and relationships visualized via Protégé Ontograf.

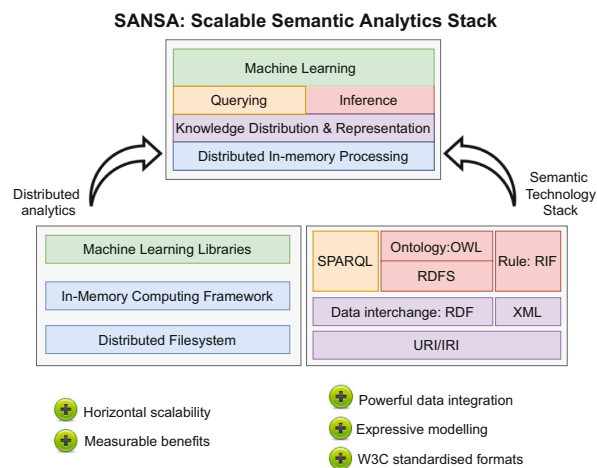


FIGURE 3 SANSa framework architecture.

In our framework, we leverage several SANSa's functionalities. We use the "Read/Write RDF library" to read the RDF data obtained from the raw data. This library allows us to read the RDF data from HDFS or the local drive file in the form of triples, and represent it in Spark's native distributed data structure, RDDs (resilient distributed datasets). To infer new information from the existing triples, we use an extended version of the "inference library" provided by SANSa. This library supports Jena and Web Ontology Language (OWL) API interfaces for processing RDF and OWL data, respectively. As both RDF and OWL contain schema information and links between different resources, applying rules enables us to infer new knowledge and expand upon the existing one. SANSa provides an adaptive rule engine that can utilize a given set of rules and derive an efficient execution plan. Finally, we use the "querying library," which provides methods for performing queries directly within programs instead of writing the code corresponding to those queries.⁶² For querying the data, we use SparQL, which allows users to query RDF graphs by specifying "templates" against which to compare graph components. Data that matches or "satisfies" a template is returned from the query.⁶³

Inference

As mentioned earlier, SANSa's inference library provides several sets of rules, which can be used to infer new knowledge from existing facts. In our implementation, we use the OWL-Horst reasoner, which contains a set of useful rules from the OWL language.⁶⁴ An example of a rule can be seen in Table 1. In this rule, if a *property*₁ is defined as "inverse" of

TABLE 1 OWL Horst rule example.

| Condition | Consequence |
|--|--|
| $property_1 \text{ owl:inverseOf } property_2$ | $instance_2 \text{ prop}_2 \text{ instance}_1$ |
| $instance_1 \text{ prop}_1 \text{ instance}_2$ | |

another $property_2$, and then we have an $instance_1$ connected through $property_1$ with an $instance_2$, the reasoner will create the relationship between $instance_2$ and $instance_1$ through $property_2$.

We can see an example of this property applied to our ontology: we have two relationships, “has” and “from,” defined as inverse. If we have a connection between an instance of the “user” class and an instance of the “game session” class through property “has” (i.e., a user has a game session), the reasoner will create the relationship between “game session” and “user” through the property “from” (i.e., game session from a user). In addition to the predefined set of rules we use for the inference, we extend this inference by using the “Triple” class implementation provided by SANSa. Our implementation allows for custom inferences, iterating over the triples, and creating or deleting new ones if necessary. This way, we can create custom rules that extend the existing ones in SANSa, adapting the inference process to our specific needs.

Querying

Querying an RDF graph is the primary method for searching, exploring, and extracting information from the underlying RDF data.¹⁹ For querying our data, we use SparQL, which is the standard language for querying RDF data. SparQL queries consist of three different parts: the pattern matching part, which includes various features of graph pattern matching, such as optional parts, the union of patterns, or nesting; the solution modifiers, which allow modifying the query results by applying operators like projection, distinct, or limit; and the output, which can be in different forms, such as yes/no or a selection of values.⁶⁵ To perform SparQL queries, SANSa implements *Sparklify*, a scalable software component for efficient evaluation of SparQL queries over distributed RDF datasets. It uses a SparQL-to-SQL rewriter to translate SparQL queries into Spark executable code.⁶⁶ Thus, our metrics are developed in the form of SparQL queries, which are executed over the RDF graph created. The use of our ontology, along with the SANSa framework and SparQL queries, enables metric interoperability, satisfying R3. To select the metrics to implement, we focused on replicating metrics from the current state of the art in GBA. For this purpose, we used the selection of papers from a previous systematic literature review in the GBA area,⁵⁵ carefully reviewing each paper and selecting the described metrics. We excluded calculations over data that involved ML, deep learning (DL), and similar models/algorithms, as our focus was on metrics described in the literature.

3.2.3 | Metrics output

When querying in SANSa, SparQL takes the description as a query and returns that information as a set of bindings or an RDF graph. In our framework, we extended this functionality, providing three different output formats:

- **Text formatting:** Query results are transformed into a text readable format, which can be stored as a text file or shown via the console output.
- **CSV formatting:** Query results are transformed into CSV format.
- **Database store:** Query results are saved into a MySQL database, which enables metric persistence for later retrieval from different applications.

3.2.4 | Authentication and authorization

We have implemented both authentication and authorization processes in our framework, addressing the issues presented in R5. Authentication is the process of identifying an entity (users) and is a prerequisite for authorization. Authorization, or access control, is the process of determining whether an entity (a device or a user) can access specific resources.⁶⁷ Specifically, we have implemented RBAC using the functionalities of the play framework.⁶⁸ This allows us to restrict access

to specific resources based on the user's role. In our framework, each user can log in to the system using a username and password, and a user can have one of the following three roles:

- **Admin role.** Users with this role can access the entire system. They have the ability to add or remove new users, insert new GBA data, and query metrics from any game and group.
- **Instructor role.** Instructors can insert new GBA data and query metrics from games and groups in which they participate.
- **Learner role.** Learners are only allowed to query their own metric results.

This way, we restrict access to different groups and games data to ensure that only appropriate users can have access. To make that possible, the system keeps a record of the games and groups related to each user (which will be the ones to which the user has access).

3.2.5 | Service API

Our service API has been developed using play framework. This scala-based solution offers an HTTP-focused framework with numerous helpers to accelerate development, resulting in shorter iterations and faster deployments. The API supports two types of calls: retrieval calls, represented by HTTP GET methods (typically used to retrieve data from a server at the specified resource), and insertion calls, represented by HTTP POST methods (used to send data to the API server to create or update a resource).

Specifically, GET methods allow users to access metrics data. Generally, these methods have the following route pattern: `/api/metricName/game/group/user`. By specifying the name of the metric, game, group, or user, the corresponding data can be accessed. Before each call, the authorization and authentication module checks that the user authenticated has the appropriate permissions to access the requested resource. On the other hand, POST methods enable users to insert data related to new users (if the user has an admin role), games, and GBA data. All information should be sent in JSON format using the body of the HTTP POST method. Moreover, when inserting data, we have implemented two different possibilities (calls), each one thought for a different purpose:

- `/api/event/addAll`: This call has been designed to process whole datasets containing a large number of events. The system will process the dataset provided as soon as possible.
- `/api/event/add`: This call has been designed for streaming-oriented systems (e.g., students are playing the game in the classroom, and the system sends log data in real-time). Thus, the events will be sent individually; the system will save each event and process the whole dataset periodically.

To ensure that the system only processes new data, each time a dataset is received, the system checks the number of events associated with each user. Only users with new events will be considered for further processing.

APIs expose data and services that consumers want to use. An API should be designed with an interface the consumer can understand, and API documentation is key to the app developers comprehending the API. For documenting our API, we have used Swagger,⁶⁹ one of the most popular API documentation frameworks. It provides a standard, language-agnostic way of defining a REST API interface, allowing the client to understand the capabilities of the REST service without any prior access to the service implementation code or network inspection.⁷⁰ The complete API specification in Swagger can be found in Reference 59. This fulfills the requirements presented in R4.

4 | PERFORMANCE EVALUATION AND CASE STUDY VALIDATION

4.1 | GBA selected metrics

As previously mentioned, we thoroughly reviewed a previous systematic literature review on GBA⁵⁵ to identify and replicate metrics implemented in previous studies. After reviewing all the metrics, we selected the following six groups of metrics:

- **Levels of activity:** This metric is computed for each game, group, and user. It includes straightforward metrics to compute based on a feature engineering process, such as the active time, inactive time, number of events, and the number of distinct types of events.
- **Persistence indicators:** This metric is computed for each game, group, and user. It includes the total amount of time spent in units (levels), the number of units completed, and the maximum time spent in a single unit.
- **Action indicators:** This metric is computed for each game, group, and user. It includes the total amount of time spent in the game and the frequency of events (number of events/total time).
- **Event types:** This metric is computed for each user and game, and it includes the number of events of each user grouped by event type (e.g., “Complete,” “Retry,” or “Interaction”). In addition, this metric group also includes the interaction level, defined as interaction events divided by the sum of the rest of the events.
- **Funnel by user:** This metric is computed for each game, group, and user. It includes the percentage of units started, the percentage of units interacted with, and the percentage of units completed by the user. This funnel provides a quick overview of each user and the game’s current status and progress.
- **User performance:** This metric is computed for each game, group, and user. It includes the percentage of success (defined as the number of units completed divided by the number of units started) and the maximum unit reached by the player.

These metrics have been implemented in our framework using SparQL queries. They are later used to test the system’s performance and serve as an example in our use case validation.

4.2 | Performance evaluation

In our performance evaluation, we evaluate the impact of our framework computation and analyze our approach’s scalability when the dataset size increases. Specifically, we focus on examining the flexibility (how quickly our approach processes different types of metrics) and scalability (how well our framework scales with larger RDF datasets). In the following subsections, we present the server configuration settings, the datasets used, and our findings.

4.2.1 | Experimental setup

For our experiments, we aimed to test our framework with real data. Therefore, we selected a diverse set of SGs from various knowledge domains to evaluate the interoperability of our approach. Field Day⁷¹ is a research lab at the Wisconsin Center for Education Research, University of Wisconsin-Madison, that designs learning games and makes their game data publicly available. From this open game data, we selected five different SGs to use their data and test the capabilities of our framework.

As we see in Table 2, each one of these datasets contains a total of 2M game events derived from real players’ interaction with the games. To test the system’s scalability when increasing the size of the dataset, we partitioned each dataset into smaller parts to have 100k, 250k, 500k, and 1M events datasets. The number of triples in our experiments varies from 1.6 to 34.8M, depending on the game and the dataset size.

TABLE 2 Dataset sizes.

| Game | Size (GB) | # of triples |
|--------------|-----------|--------------|
| Crystal | 2.98 | 34,500,365 |
| Balloon | 2.87 | 32,665,037 |
| Cycle carbon | 2.93 | 33,973,329 |
| Magnet | 2.76 | 31,249,623 |
| Waves | 3.01 | 34,780,117 |

We implemented our approach using Python 3.8, Spark-3.0.1, Scala 2.12.11, Java 11, and all the data were stored on an HDFS cluster using Hadoop 2.10.2. All experiments were conducted on a cluster of six nodes: one master and five workers. The cluster had a total of 36 cores (six cores per worker), 112 GB RAM (32 GB for the server node, 16 GB for each worker), and 3 TB SSD storage with a speed of 12 GB/s.

4.2.2 | Performance results

We evaluate our approach using the experimental setup described in the previous section and the metrics described in Section 4.1. We assessed the runtime of our distributed framework throughout the entire pipeline, from processing the raw events to calculating the metrics using SparQL queries. We ran experiments on five different sizes to measure the performance of size-up scalability. Since the ontology data processing stage is run locally, we did not include this as part of the node scalability performance evaluation. The average execution time of this stage is shown in Table 3. Then, to measure the performance of node scalability, we run experiments using one to five worker nodes on each of the five dataset sizes. Since we selected data from five games, we executed each experiment using those five datasets. The average execution time is presented in Table 4 and Figure 4.

In Table 4, we highlight the best execution time for each dataset size and stage/query in green. As we can see, the inference time benefits from increasing the number of workers as the dataset size increases. In fact, we see that the best execution time for 100k events is given by using two workers; meanwhile, the best execution time for 2M events is given by using five workers. Regarding the different metrics, querying the RDF triples also benefits from increasing the number of workers. However, in most cases, the best performance is achieved using four workers, except for some specific metrics with 1 and 2M events, which show better performance using five workers. In addition, we also see that the one worker cluster fails to process the 2M events dataset due to working memory errors.

As we can observe in Figure 5, the execution time grows linearly when the size of the datasets increases, demonstrating the scalability of our approach when using three or more workers. Furthermore, the query execution time varies depending on the metric being computed. For instance, with a dataset of 2M events, the framework can compute the “action indicators” metric in an average of 13.6 s, while the “event types” metric takes an average of 203.8 s to calculate. This discrepancy is due to the different SparQL queries designed for each metric, as they involve different types of operations. Taking the same example as before, the “action indicators” metric uses simple filtering and aggregation operations. In contrast, the “event types” metric uses several join operations, significantly increasing the computational cost.

The total execution time for each experiment is shown in Figure 4. When using smaller datasets, we can see that the performance improvement when using more workers could be more remarkable. However, with four and five workers, performance slightly decreases. For instance, when computing a 100k events file, the average execution time increases from 180.6 s using two workers to 199.8 s using five workers. With larger datasets (1 and 2M events), there is a performance improvement when using three and four workers, but the impact of using five workers is not significant. For the 1M events experiments, we obtain an average execution time of 1047.8 s using four workers and an average of 1070.6 s using five workers. When computing 2M events, if we compare the two workers configuration and the four workers configuration, the average execution time is 69% lower using four workers. With these results, combined with the fact that most of the lowest query execution times were given by the four workers cluster, we can affirm that using this configuration is the best option to obtain better performance and save the resources required by an additional worker node.

TABLE 3 Ontology data processing execution time.

| # Events | Ontology data processing time (s) (mean) |
|----------|--|
| 100k | 80.2 |
| 250k | 271 |
| 500k | 762.4 |
| 1M | 2112.8 |
| 2M | 4017 |

TABLE 4 Performance analysis on large-scale GBA datasets.

| # Events | # Workers | Runtime (s) (mean) | | | | | | |
|----------|-----------|--------------------|--------------------|-------------|---------------------|-------------|----------------|------------------|
| | | Inference | Levels of activity | Persistence | Activity indicators | Event types | Funnel by user | User performance |
| 100k | 1 | 92 | 20 | 16.2 | 8.2 | 51.6 | 22.4 | 10.6 |
| | 2 | 83.4 | 15 | 10.6 | 6 | 41.6 | 16.8 | 7.2 |
| | 3 | 86.4 | 15 | 10.6 | 6 | 41.6 | 16.8 | 7.2 |
| | 4 | 97.4 | 13.8 | 8.4 | 5.2 | 37.4 | 16.2 | 5 |
| | 5 | 103.2 | 15.8 | 10.2 | 5.2 | 40.4 | 17.8 | 7.2 |
| 250k | 1 | 287.4 | 27.4 | 20.8 | 11.6 | 71.8 | 34.6 | 14.8 |
| | 2 | 216.8 | 16.6 | 12.4 | 7.2 | 49.6 | 26 | 9.6 |
| | 3 | 202 | 15.2 | 9.8 | 6.4 | 44.4 | 23.6 | 7.6 |
| | 4 | 223.4 | 15 | 9 | 5.4 | 41.4 | 23.8 | 7.2 |
| | 5 | 248 | 17.6 | 12.6 | 7 | 44.6 | 29.8 | 8.6 |
| 500k | 1 | 744.8 | 37.8 | 26 | 13.2 | 110 | 50 | 21.2 |
| | 2 | 503 | 27.6 | 15.8 | 9 | 70.6 | 34.8 | 11.6 |
| | 3 | 404.6 | 24.2 | 13.8 | 9.2 | 66.2 | 41.2 | 10 |
| | 4 | 437 | 23 | 10.4 | 9.6 | 54.2 | 38 | 8.6 |
| | 5 | 433.8 | 23.2 | 13.2 | 10.6 | 57.6 | 43 | 9.8 |
| 1M | 1 | 1249.8 | 60 | 33.4 | 19.2 | 218 | 78.6 | 30.8 |
| | 2 | 902.2 | 38 | 20 | 14.2 | 151 | 47.8 | 16 |
| | 3 | 812.4 | 31.4 | 18.6 | 16.2 | 149.8 | 45.2 | 15.6 |
| | 4 | 788.8 | 33.4 | 17.2 | 11.6 | 144 | 41.6 | 11.2 |
| | 5 | 807 | 32.6 | 18.4 | 11.4 | 149.4 | 40.8 | 11 |
| 2M | 1 | FAIL | | | | | | |
| | 2 | 2803 | 63.8 | 33.2 | 16.4 | 231.6 | 140.6 | 25.6 |
| | 3 | 1940.6 | 55.8 | 28.2 | 14.8 | 218.6 | 132.2 | 22 |
| | 4 | 1532.8 | 51.8 | 24.2 | 13.6 | 203.8 | 123.2 | 16.8 |
| | 5 | 1522.6 | 56.6 | 29.6 | 14.6 | 215 | 119.4 | 19 |

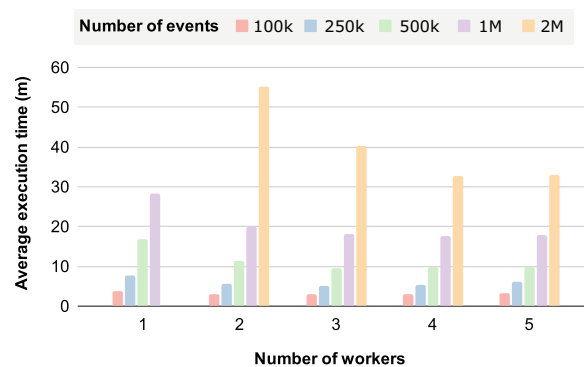


FIGURE 4 Size and worker scalability analysis.

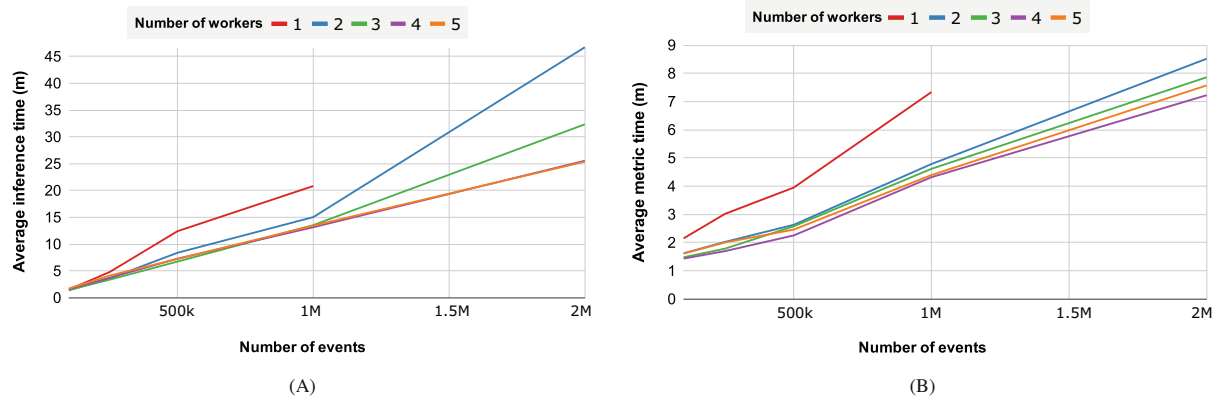


FIGURE 5 Scalability analysis. (A) Inference execution time per number of events and workers. (B) Metric execution time per number of events and workers.

4.3 | Case study validation

In this section, we present a case study with two use cases to exemplify how our framework and metrics can be applied in a real context. First, we conducted a use case by designing and implementing a dashboard that utilizes the analyzed and transformed data in the form of metrics, which can be consumed through visualizations. Second, we implemented a learner report system that enables instructors to easily track their learners' progress over time. The use cases performed show how, using the API, we provide a straightforward interface that can be used from almost any device, meeting **R4** and using the results from meeting the rest of **Requirements**.

4.3.1 | Use case: Dashboard

In this first use case, we introduce a visualization dashboard system that leverages the data analyzed and transformed into metrics by our framework. This dashboard utilizes specific API calls to input new data into the system or retrieve existing metrics from different games and groups. It enables (1) instructors to monitor learners' interactions with games, adapting their interventions based on these insights or using the metrics for formative evaluation, and (2) learners to track their own game-related activity. The dashboard aligns with the different roles defined within the framework. We developed the dashboard using the Shiny framework in R and deployed it on the ShinyApps web server.

In Figure 6, we can see the dashboard running live on the ShinyApps server. Users can log into the system using a username and a password. Each user will have different permissions and functionalities depending on the credentials used, complying with **R5**. For example, instructors and administrators can upload new GBA data in the "file upload" tab. Users can also navigate through the available tabs to upload new data or query the different metric results calculated.

Finally, we can see how the dashboard fully benefits the interoperability between games and metrics. The user can use selection boxes to choose between games and groups, and also between users depending on the granularity of the metric. That way, when a game is selected among the available options, the system loads the existing groups for that specific game in the corresponding selection box. When a group is selected, the system loads the existing users for that specific group. Once all the selection boxes for that metric tab are filled with a choice, the system queries the necessary information and represents it using interactive visualizations, as shown in Figure 6A,B.

4.3.2 | Use case: Reports

This second use case consists of a learner report summarizing the learners' progress using different games. The report is automatically generated using RMarkdown and outputs a PDF or HTML file that is sent to instructors periodically (the frequency in which the reports are generated can be adjusted in the system). Regarding the connection with the

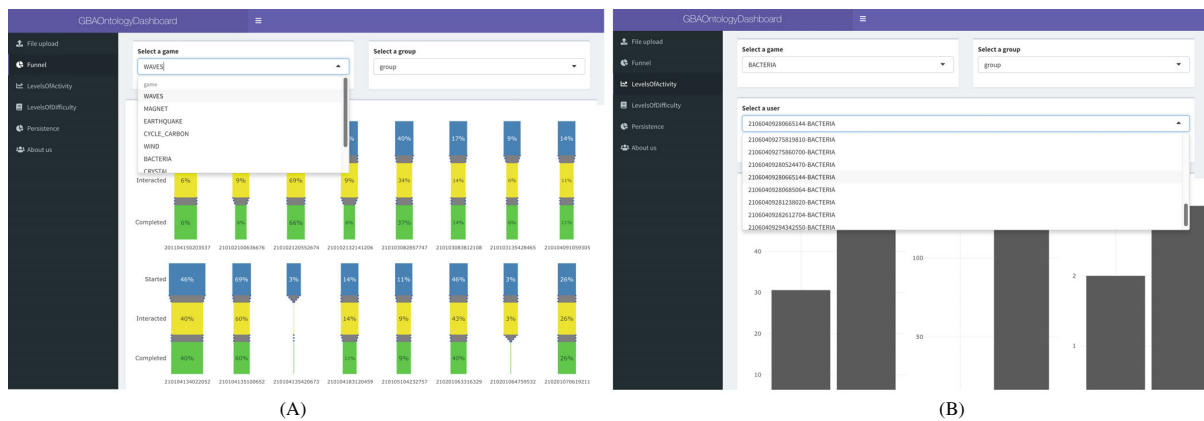


FIGURE 6 Screenshots of the dashboard developed. (A) Game and group selection box in funnel by puzzle metric. (B) Game, group and user selection box in levels of activity metric.

framework, our report system uses specific query API calls, and the retrieved results are then employed to build the different parts of the report. For this specific example, we selected an instructor that has access to two different games: EARTHQUAKE and MAGNET, and specifically to the “MainGroup” data of each game. Some examples of information that these reports include are shown in Figure 7. Specifically, in Figure 7A we can see the first part of the report. Here, we see a group summary for each game, including some key metrics such as the total active time in seconds, the total number of units started, the total number of units completed, or the units that have been more problematic for learners. Then, in Figure 7B, we can see an individual report for each learner, also showing similar key metrics; and finally, in Figure 7C, a plot showing the learners’ persistence. In this plot, each bubble represents a different learner, the x-axis represents the number of attempted units, and the size of each bubble represents the number of completed units. Moreover, the y-axis represents the average persistence percentile, so more persistent learners will be at the top of the plot. This report provides an easy way to monitor groups and learners while playing different games, allowing instructors to perform quick assessments based on different metrics calculated automatically using learners’ data.

5 | DISCUSSION

SGs are considered practical tools in multiple domains. In particular, it is believed that its use for assessment (GBA) will be an increasing part of testing programs in future generations. This is due to their promising possibilities for more valid and reliable measurement of learners’ skills compared to traditional assessment methods, such as paper-and-pencil tests.⁷² However, the time and cost-intensive process of developing digital learning or assessment environments restricts the practical implementation of GBAs. Furthermore, the limited interoperability of assessment and tracking systems across different platforms presents a critical constraint in this area.⁷³ Our approach addresses these limitations by developing a framework that incorporates an intermediate semantic layer to enable interoperable GBAs. By utilizing an ontology as a common knowledge model, our framework can integrate log events from diverse games into a unified data model. This, combined with the use of interoperable RDF metrics, promotes standardization in the field and facilitates the utilization of numerous games, each designed for specific purposes, knowledge domains, and target participants.

Another known constraint in the area is the use of small data sample sizes. Generally, sample sizes used in GBA studies are pretty limited in size, resulting in low statistical power and a reduced chance of detecting actual effects.^{53,74} Although collecting large samples of in-context data is a challenging task,⁵⁵ future research should use larger data samples in order to improve the results generalization and validity. This would also enable the use of more complex techniques, such as neural networks, which often require large amounts of data to outperform other models. Our contribution involves leveraging big data technologies to efficiently process large quantities of GBA data. Using a cluster of four worker nodes, our framework can process 2M events (including the computation of the six different metrics) in an average of 6434 s (107.2 min). We estimated that each user produces approximately 512 game events/hour based on fifteen different datasets from our experiments. Considering a classroom of 25 learners using a game for one hour/week, it would result in 51,200

GBA Ontology Framework Report

This report has been generated automatically for instructor with username: teacher1

This report includes MainGroup from the game EARTHQUAKE and MainGroup from the game MAGNET.

Group summary

This has been the performance of each group for the selected games:

Summary for each group.

| Game | Group | Avg Active Time By Attempt | Total Active Time | Avg Events By Attempt | Total Events | Total Units Started | Total Units Completed | Most Difficult Unit |
|------------|-----------|----------------------------|-------------------|-----------------------|--------------|---------------------|-----------------------|---------------------|
| EARTHQUAKE | MainGroup | 108.82055 | 21398.25 | 21.954315 | 4325 | 354 | 16 | EARTHQUAKE-36 |
| MAGNET | MainGroup | 34.69733 | 157005.42 | 8.273149 | 37436 | 5935 | 3318 | MAGNET-58 |

For each game, the units that could be problematic (abandoned percentage > 75%) are:

| Game | Unit | Percentage Abandoned |
|------------|---------------|----------------------|
| EARTHQUAKE | EARTHQUAKE-18 | 100.0000 |
| EARTHQUAKE | EARTHQUAKE-19 | 100.0000 |
| EARTHQUAKE | EARTHQUAKE-36 | 76.4706 |
| EARTHQUAKE | EARTHQUAKE-47 | 100.0000 |
| EARTHQUAKE | EARTHQUAKE-6 | 100.0000 |
| MAGNET | MAGNET-0 | 100.0000 |

(A)

General student report

This section provides a general overview of users' progress in selected games.

General performance

Summary for each user.

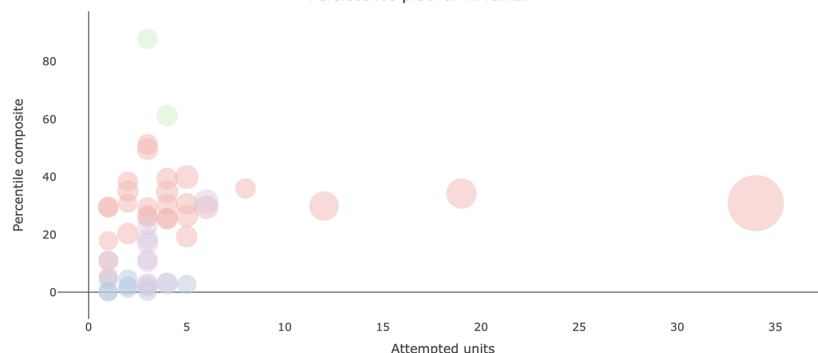
| Game | User | Avg Active Time By Attempt | Total Active Time | Total Inactive Time | Avg Events By Attempt | Total Events | Total Units Started | Total Units Completed | Total Time |
|------------|-------------------|----------------------------|-------------------|---------------------|-----------------------|--------------|---------------------|-----------------------|------------|
| EARTHQUAKE | 21050212144374524 | 127.9285 | 255.857 | 113.035 | 21.5 | 43 | 2 | 0 | 368.892 |
| EARTHQUAKE | 21050408083960244 | 79.2530 | 158.506 | 124.032 | 18.0 | 36 | 2 | 0 | 282.538 |
| EARTHQUAKE | 21050410373924384 | 61.8830 | 61.883 | 79.698 | 34.0 | 34 | 1 | 0 | 141.581 |
| EARTHQUAKE | 21060113472500124 | 10.4990 | 10.499 | 0.000 | 1.0 | 1 | 2 | 0 | 10.499 |
| EARTHQUAKE | 21070005462090520 | 151.8510 | 151.851 | 40.175 | 50.0 | 50 | 1 | 0 | 192.026 |
| EARTHQUAKE | 21070008000370016 | 8.3100 | 8.310 | 0.000 | 1.0 | 1 | 1 | 0 | 8.310 |
| EARTHQUAKE | 21070008572818384 | 101.0990 | 101.099 | 308.638 | 36.0 | 36 | 1 | 0 | 409.737 |
| EARTHQUAKE | 21070009344713380 | 77.0570 | 77.057 | 0.000 | 5.0 | 5 | 1 | 0 | 77.057 |
| EARTHQUAKE | 21070012163629812 | 63.7750 | 63.775 | 184.488 | 15.0 | 15 | 1 | 0 | 248.263 |
| EARTHQUAKE | 21070012553624696 | 50.6720 | 50.672 | 0.000 | 8.0 | 8 | 1 | 0 | 50.672 |

(B)

Persistence

The persistence summary for each one of the games has been:

Persistence plot for MAGNET



Note that the size of each bubble represents the number of completed units for each user.

(C)

FIGURE 7 Learners' report screenshots. (A) Group summary. (B) General student report. (C) General learners report.

events per class and month. This implies that our framework can process data from approximately 39 full classrooms for an entire month in just 107.2 min. Additionally, our approach supports streaming data, as log events are received individually, allowing for real-time processing and just-in-time feedback. Since only new events are considered in each processing iteration, the data size processed is significantly reduced.

In-game metrics are necessary and essential, but we have to choose the most appropriate ones depending on each project. The most used metrics everywhere in any platform are performance metrics.⁷⁵ However, beyond performance, we can obtain further insights from the analysis of learner-generated information. Actions and behaviors should be convertible into metrics to identify learners' individual characteristics (including behaviors, performance, or skills) and learner-generated game data (e.g., time spent, goals, tasks completed).^{52,76} After reviewing previous GBA literature, we found a set of commonly used metrics to replicate in our system. As mentioned earlier, one challenge is that these metrics and indicators are typically designed and developed specifically for each game. In this study, we have successfully replicated and integrated all of these previously established metrics into our framework, showing the interoperability between different games and excellent performance using large-scale datasets. For example, the "levels of activity" metric takes an average of 51.8 s to compute using 2M events, and the metric "user performance" takes an average of only 16.8 s using the same number of events. This achievement is made possible through the querying module, which translates SparQL code into Spark executable code, enabling the creation of interoperable metrics that measure not only performance-related characteristics but also other types of skills and behaviors. For example, we could apply clustering techniques to identify

distinct student behaviors by utilizing SparQL code to collect student features, followed by employing the ML module provided by SANSA.

Typically, GBA systems develop their own interfaces to interact with external data sources. In our system, we have integrated a service API, which allows for easy insertion and retrieval of data, facilitating the interaction of various sources with our framework. One of the main advantages of our approach is its simplicity and ease of use. By integrating the API, users can easily build applications that connect to the framework and access its capabilities. This approach offers a number of benefits, including the ability to scale the service to meet the demands of a large number of users, the ability to easily update and maintain the service, and the ability to offer a seamless user experience across different devices. Overall, the integration of an API into a framework is a key enabler for the GBaaS paradigm and offers a number of advantages for both GBA researchers and users. One potential limitation is the risk of external users accessing confidential information. To address this concern, we have developed an authorization and authentication module that controls access to each resource, ensuring maximum user data privacy.

This work also has some limitations: first, although we have defined an ontology with terms and concepts that almost any log data from the area should have, there is still a manual process of adapting the GBA data to our ontology to meet the input's requirements. Future researchers could take into account the ontology in the collected data design to skip this manual step. Second, the ontology processing data stage is run locally, which does not allow to take full advantage of the possibilities and performance that distributed-systems have. In addition, our framework (with the current configuration described in Section 4.2.1) cannot compute datasets with more than 3M events in a single batch due to working memory limitations. However, the system can solve this increasing resources or by splitting those files into smaller chunks and processing them sequentially. Finally, the system supports ML techniques but does not support more complex methods, such as knowledge inference or DL, which are also common in the GBA field.⁵² Using these methods could help infer more helpful information from learners' data and improve the results' validity and reliability.

6 | CONCLUSIONS AND FUTURE WORK

This research aimed to create a robust novel framework for enabling GBaaS using ontologies and big data technologies. Moreover, we demonstrated its capabilities by replicating existing metrics in GBA literature and conducting a case study with two use cases to show how external users can consume the system as a service. We also conducted a performance evaluation using different cluster configurations, concluding that using a cluster of one master node and four worker nodes was the best option in terms of resource management and performance. This cluster configuration was capable of processing 2M user events (approximately the size of 39 classrooms using a game for one hour/week for one month) in an average of 107.2 min.

As part of our future work, we want to validate our approach by conducting case studies in which the framework will be used in real-time, collecting data from learners and instructors, and validating the data streaming functionality implemented. Moreover, we would like to continue developing new GBA metrics that could use more advanced techniques, such as ML algorithms. Additionally, DL models could also be developed to test their predictive performance for inferring students' knowledge using existing data. Despite all the benefits that the application of ML and DL could have, most non-technical users perceive them as "black boxes." In this regard, future work should address the use of eXplainable Artificial Intelligence (XAI) approaches, enabling non-technical users (such as teachers) to interpret AI-generated insights and recommendations, empowering them to make informed decisions. Finally, we plan to integrate the ontology data processing stage (which is currently running locally) into the distributed environment to take full advantage of the cluster capabilities and obtain even better performance results. Future work could also address the deployment of the service in the cloud. Data and applications hosted on the cloud allow businesses to be more responsive and adaptable, becoming more efficient, strategic, and insight-driven.⁷⁷ Additionally, the use of fog computing could also be introduced in our platform, extending cloud computing due to its low latency, energy efficiency and the reduction in bandwidth required for data transport.⁷⁷

This research contributes significantly to the current state of the art, including a completely novel framework that enables interoperable GBAs using large-scale data, privacy management, and easy interaction from external sources. We expect our contributions to solve current limitations regarding GBA interoperability, reducing the cost and effort that designing and performing specific GBAs have and allowing the deployment of GBaaS.

AUTHOR CONTRIBUTIONS

Manuel J. Gomez: Conceptualization; methodology; software; validation; formal analysis; data curation; writing – original draft; visualization. **José A. Ruipérez-Valiente:** Conceptualization; writing – review and editing; supervision; project administration. **Félix J. García Clemente:** Conceptualization; writing – review and editing; supervision; project administration.

ACKNOWLEDGMENTS

This work was partially supported by (a) Grant 21795/FPI/22 - Séneca Foundation. Cofinanced by Innovatiio Global Educación. Region of Murcia (Spain), and (b) REASSESS project (Grant 21948/JLI/22), funded by the Call for Projects to Generate New Scientific Leadership, included in the Regional Program for the Promotion of Scientific and Technical Excellence Research (2022 Action Plan) of the Seneca Foundation, Science and Technology Agency of the Region of Murcia.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Manuel J. Gomez  <https://orcid.org/0000-0003-0571-2923>

José A. Ruipérez-Valiente  <https://orcid.org/0000-0002-2304-6365>

Félix J. García Clemente  <https://orcid.org/0000-0001-6181-5033>

REFERENCES

1. Gros B. Digital games in education: the design of games-based learning environments. *J Res Technol Educ*. 2007;40(1):23-38.
2. Daniel M, Garry C. *Video Games as Culture: Considering the Role and Importance of Video Games in Contemporary Society*. Routledge; 2018.
3. Marklund BB, Backlund P, Engstrom H. The practicalities of educational games: challenges of taking games into formal educational settings. *2014 6th International Conference on Games and Virtual Worlds for Serious Applications*. IEEE; 2014:1-8.
4. Boyle EA, Hainey T, Connolly TM, et al. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Comput Educ*. 2016;94:178-192.
5. Alonso-Fernández C, Cano AR, Calvo-Morata A, Freire M, Martínez-Ortiz I, Fernández-Manjón B. Lessons learned applying learning analytics to assess serious games. *Comput Human Behav*. 2019;99:301-309.
6. Laamarti F, Eid M, El Saddik A. An overview of serious games. *Int J Comput Games Technol*. 2014;2014:358152.
7. Kato PM, Klerk S. Serious games for assessment: welcome to the jungle. *J Appl Test Technol*. 2017;18(S1):1-6.
8. Shute VJ. Stealth assessment in computer-based games to support learning. *Comput Games Instr*. 2011;55(2):503-524.
9. Kim YJ, Shute VJ. The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Comput Educ*. 2015;87:340-356.
10. Ruipérez-Valiente JA, Gomez MJ, Martínez PA, Kim YJ. Ideating and developing a visualization dashboard to support teachers using educational games in the classroom. *IEEE Access*. 2021;9:83467-83481.
11. Fathy N, Gad W, Badr N. A unified access to heterogeneous big data through ontology-based semantic integration. *2019 Ninth International Conference on Intelligent Computing and Information Systems*. IEEE; 2019:387-392.
12. Gómez-Pérez A, Fernández-López M, Corcho O. *Ontological Engineering: with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media; 2006.
13. Ifenthaler D, Eseryel D, Ge X. *Assessment for Game-Based Learning*. Springer; 2012:1-8.
14. Freire M, Serrano-Laguna Á, Manero B, Martínez-Ortiz I, Moreno-Ger P, Fernández-Manjón B. Game learning analytics: learning analytics for serious games. In: Spector M, Lockee B, Childress M, eds. *Learning, Design, and Technology*. Springer; 2016:1-29.
15. Bienkowski M, Feng M, Means B. *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: an Issue Brief*. US Department of Education: Office of Educational Technology; 2012.
16. Serrano Á, Marchiori EJ, Blanco Á, Torrente J, Fernández-Manjón B. A framework to improve evaluation in educational games. *Proceedings of the 2012 IEEE Global Engineering Education Conference*. IEEE; 2012:1-8.
17. Sejdiu G, Rula A, Lehmann J, Jabeen H. A scalable framework for quality assessment of RDF datasets. In: Ghidini C, Hartig O, Maleshkova M, et al., eds. *The Semantic Web – ISWC 2019*. Springer; 2019:261-276.
18. Mami MN, Graux D, Scerri S, Jabeen H, Auer S, Lehmann J. Squerall: virtual ontology-based access to heterogeneous and large data sources. In: Ghidini C, Hartig O, Maleshkova M, et al., eds. *The Semantic Web – ISWC 2019*. Springer; 2019:229-245.

19. Lehmann J, Sejdiu G, Bühmann L, et al. Distributed semantic analytics using the SANSa stack. *International Semantic Web Conference*. Springer; 2017:147-155.
20. Susi T, Johannesson M, Backlund P. Serious games: an overview. Technical Report. Institutionen för kommunikation och information; 2007.
21. Gomez MJ, Ruipérez-Valiente JA, Martínez PA, Kim YJ. Applying learning Analytics to detect sequences of actions and common errors in a geometry game. *Sensors*. 2021;21(4):1025.
22. Wang R, DeMaria S Jr, Goldberg A, Katz D. A systematic review of serious games in training health care professionals. *Simul Healthc*. 2016;11(1):41-51.
23. Proença JP, Quaresma C, Vieira P. Serious games for upper limb rehabilitation: a systematic review. *Disabil Rehabil Assist Technol*. 2018;13(1):95-100.
24. Samčović AB. Serious games in military applications. *Vojnoteh Glas*. 2018;66(3):597-613.
25. Albaladejo-González M, Strukova S, Ruipérez-Valiente JA, Gómez Mármol F. Exploring the affordances of multimodal data to improve cybersecurity training with cyber range environments. VI Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2021); 2021.
26. Ruipérez-Valiente JA, Gaydos M, Rosenheck L, Kim YJ, Klopfer E. Patterns of engagement in an educational massively multiplayer online game: a multidimensional view. *IEEE Trans Learn Technol*. 2020;13(4):648-661.
27. Harpstead E, Zimmermann T, Nagapan N, et al. What drives people: creating engagement profiles of players from game log data. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM; 2015:369-379.
28. DiCerbo KE. Game-based assessment of persistence. *J Educ Technol Soc*. 2014;17(1):17-28.
29. Ruipérez-Valiente JA, Kim YJ. Effects of solo vs. collaborative play in a digital learning game on geometry: results from a K12 experiment. *Comput Educ*. 2020;159:104008.
30. Kim YJ, Lin G, Ruipérez-Valiente JA. *Expanding Teacher Assessment Literacy with the Use of Data Visualizations in Game-Based Assessment*. Springer; 2021:399-419.
31. Jaffal Y, Wloka D. Employing game analytics techniques in the psychometric measurement of game-based assessments with dynamic content. *J e-Learn Knowl Soc*. 2015;11(3):101-115.
32. Dziejczak D, Włodarczyk W. Approaches to measuring the difficulty of games in dynamic difficulty adjustment systems. *Int J Hum-Comput Interact*. 2018;34(8):707-715.
33. Kiili K, Moeller K, Ninaus M. Evaluating the effectiveness of a game-based rational number training-in-game metrics as learning indicators. *Comput Educ*. 2018;120:13-28.
34. Lindenmayer JP, Goldring A, Borne S, et al. Assessing instrumental activities of daily living (iADL) with a game-based assessment for individuals with schizophrenia. *Schizophr Res*. 2020;223:166-172.
35. Jackson DJ, Kim S, Lee C, Choi Y, Song J. Simulating déjà vu: what happens to game performance when controlling for situational features? *Comput Human Behav*. 2016;55:796-803.
36. Crytek U. *Crysis 2*. Electronic Arts; 2011.
37. Serrano-Laguna Á, Martínez-Ortiz I, Haag J, Regan D, Johnson A, Fernández-Manjón B. Applying standards to systematize learning analytics in serious games. *Comput Stand Interfaces*. 2017;50:116-123.
38. Said B, Cheniti-Belcadhi L, El Khayat G. An ontology for personalization in serious games for assessment. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE; 2019:148-154.
39. Rocha OR, Zucker CF. Ludo: an ontology to create linked data driven serious games. ISWC 2015—Workshop on LINKed EDucation; 2015.
40. Staab S, Studer R. *Handbook on Ontologies*. Springer Science & Business Media; 2010.
41. Zada I, Shahzad S, Ali S, Mehmood RM. OntoSuSD: Software engineering approaches integration ontology for sustainable software development. *Softw Pract Exp*. 2023;53(2):283-317.
42. Al-Chalabi HKM, Hussein AMA. Ontology applications in E-learning systems. *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE; 2020:1-6.
43. Andrews-Todd J, Kerr D. Application of ontologies for assessing collaborative problem solving skills. *Int J Test*. 2019;19(2):172-187.
44. Gomez MJ, Ruipérez-Valiente JA, Clemente FJG. Developing and Validating Interoperable Ontology-Driven Game-Based Assessments (In Review); 2022.
45. Botoeva E, Calvanese D, Cogrel B, Corman J, Xiao G. A generalized framework for ontology-based data access. In: Ghidini C, Magnini B, Passerini A, Traverso P, eds. *AI*IA 2018 – Advances in Artificial Intelligence*. Springer; 2018:166-180.
46. Zhou J, Ma L, Liu Q, Zhang L, Yu Y, Pan Y. Minerva: a scalable OWL ontology storage and inference system. In: Mizoguchi R, Shi Z, Giunchiglia F, eds. *The Semantic Web – ASWC 2006*. Springer; 2006:429-443.
47. Dehainsala H, Pierra G, Bellatreche L. Ontodb: an ontology-based database for data intensive applications. In: Kotagiri R, Krishna PR, Mohania M, Nantajeewarawat E, eds. *International Conference on Database Systems for Advanced Applications*. Springer; 2007:497-508.
48. Abbes H, Gargouri F. Big data integration: a MongoDB database and modular ontologies based approach. *Procedia Comput Sci*. 2016;96:446-455.
49. Mountasser I, Ouhbi B, Hdioud F, Frikh B. Semantic-based big data integration framework using scalable distributed ontology matching strategy. *Distrib Parallel Databases*. 2021;39(4):891-937.
50. Reyes-Álvarez L, Roldán-García MM, Aldana-Montes JF. Tool for materializing OWL ontologies in a column-oriented database. *Softw Pract Exp*. 2019;49(1):100-119.
51. Verma V, Baron T, Bansal A, Amresh A. Emerging practices in game-based assessment. In: Ifenthaler D, Kim YJ, eds. *Game-Based Assessment Revisited*. Springer; 2019:327-346.

52. Kim YJ, Ifenthaler D. Game-based assessment: the past ten years and moving forward. In: Ifenthaler D, Kim YJ, eds. *Game-Based Assessment Revisited*. Springer; 2019:3-11.
53. Alonso-Fernandez C, Calvo-Morata A, Freire M, Martinez-Ortiz I, Fernández-Manjón B. Applications of data science to game learning analytics data: a systematic literature review. *Comput Educ*. 2019;141:103612.
54. Pérez-Berenguer D, García-Molina J. A standard-based architecture to support learning interoperability: a practical experience in gamification. *Softw Pract Exp*. 2018;48(6):1238-1268.
55. Gomez MJ, Ruipérez-Valiente JA, Clemente FJG. A systematic literature review of game-based assessment studies: trends and challenges. *IEEE Trans Learn Technol*. 2022.
56. Gursoy ME, Inan A, Nergiz ME, Saygin Y. Privacy-preserving learning analytics: challenges and techniques. *IEEE Trans Learn Technol*. 2016;10(1):68-81.
57. Takabi H, Joshi JB, Ahn GJ. Security and privacy challenges in cloud computing environments. *IEEE Secur Priv*. 2010;8(6):24-31.
58. Fernández-López M, Gómez-Pérez A, Juristo N. Methontology: from ontological art towards ontological engineering. *Proceedings of the AAAI97 Spring Symposium*. AAAI Press; 1997:33-40.
59. Gomez MJ, Ruipérez-Valiente JA, García Clemente FJ. Supplementary materials: a framework to support interoperable game-based assessments as a service (GBAaaS): design, development, and use cases; 2022. Accessed December 16, 2022. https://osf.io/ctb8p/?view_only=fe268264bd1346668ae05fa2e3048f8c
60. Bischof S, Decker S, Krennwallner T, Lopes N, Polleres A. Mapping between RDF and XML with XSPARQL. *J Data Semant*. 2012;1(3):147-185.
61. Gandon F, Bottollier V, Corby O, Durville P. RDF/XML source declaration; 2007.
62. Analytics SD. Sparklify; 2021.
63. Grobe M. RDF, Jena, SparQL and the 'semantic web'. *Proceedings of the 37th Annual ACM SIGUCCS Fall Conference: Communication and Collaboration*. Association for Computing Machinery; 2009:131-138.
64. Kim JM, Park YT. Scalable OWL-Horst ontology reasoning using SPARK. *2015 International Conference on Big Data and Smart Computing*. IEEE; 2015:79-86.
65. Pérez J, Arenas M, Gutierrez C. Semantics and complexity of SPARQL. *ACM Trans Database Syst*. 2009;34(3):1-45.
66. Stadler C, Sejdin G, Graux D, Lehmann J. Sparklify: a scalable software component for efficient evaluation of SPARQL queries over distributed RDF datasets. In: Ghidini C, Hartig O, Maleshkova M, et al., eds. *The Semantic Web – ISWC 2019*. Springer; 2019:293-308.
67. Kim H, Lee EA. Authentication and authorization for the Internet of Things. *IT Prof*. 2017;19(5):27-33.
68. Hunt J. Play framework. *A Beginner's Guide to Scala, Object Orientation and Functional Programming*. Springer; 2018:431-446.
69. Software S. API development for everyone; 2022. Accessed April 28, 2022.
70. De B. *API Documentation*. Springer; 2017:59-80.
71. Day F. We're field day; 2022. Accessed April 29, 2022.
72. Klerk S, Kato PM. The future value of serious games for assessment: where do we go now? *J Appl Test Technol*. 2017;18(S1):32-37.
73. Hruska M, Long R, Amburn C, Kilcullen T, Poepelman T. Experience API and team evaluation: evolving interoperable performance assessment. *Proceedings of the Interservice/Industry Training, Simulation & Education Conference*; 2014.
74. Petri G, Wangenheim CG. How games for computing education are evaluated? A systematic literature review. *Comput Educ*. 2017;107:68-90.
75. Junaidi J, Julianto A, Anwar N, Safrizal S, Warnars HLHS, Hashimoto K. Perfecting a video game with game metrics. *TELKOMNIKA*. 2018;16(3):1324-1331.
76. Loh CS, Sheng Y. Measuring expert performance for serious games analytics: from data to insights. *Serious Games Analytics*. Springer; 2015:101-134.
77. Gill SS, Xu M, Ottaviani C, et al. AI for next generation computing: emerging trends and future directions. *Internet Things*. 2022;19:100514.

How to cite this article: Gomez MJ, Ruipérez-Valiente JA, García Clemente FJ. A framework to support interoperable Game-based Assessments as a Service (GBAaaS): Design, development, and use cases. *Softw Pract Exper*. 2023;53(11):2222-2240. doi: 10.1002/spe.3254

4 Integrating Explainable AI in Performance Prediction

Title

Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games

Authors

Manuel J. Gomez¹, Álvaro Armada Sánchez¹,
Mariano Albaladejo-González¹, Félix J. García Clemente¹,
José A. Ruipérez-Valiente¹

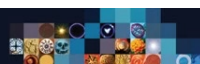
¹*Department of Information and Communications Engineering,
University of Murcia, Spain*

Publication details

| | | | |
|----------------|----------------|------------------|--------------------|
| Journal | Expert Systems | Publisher | Wiley |
| Volume | 42 | Number | 3 |
| Pages | e70008 | Year | 2025 |
| JIF | 2.3 | Rank | Q2 |
| Status | Published | DOI | 10.1111/exsy.70008 |

Abstract

In recent years, serious games (SGs) have emerged as a powerful tool in education by combining pedagogy and entertainment, facilitating the acquisition of knowledge and skills in engaging environments. SGs enable the collection of valuable interaction data from students, allowing for the analysis of student performance, with artificial intelligence (AI) playing a key role in processing this data to make informed inferences about their knowledge and skills. However, the lack of explainability in AI models represents a significant challenge. This research aims to develop an interpretable model for predicting students' performance in real-time while playing an SG by: (1) calculating the performance of an interpretable prediction model of task completion in an SG and (2) demonstrating the application of the interpretable model for just-in-time (JIT) classroom interventions. Our results show that we are able to predict students' task completion in real-time with a balanced accuracy result of 77.21% after a short playtime has elapsed. In addition, an explainable artificial intelligence (XAI) approach has been applied to ensure the interpretability of the developed models. This approach supports personalised learning experiences, unlocks AI benefits for non-technical users, and maintains transparency in education.



ORIGINAL ARTICLE

Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games

Manuel J. Gomez  | Álvaro Armada Sánchez | Mariano Albaladejo-González  | Félix J. García Clemente  | José A. Ruipérez-Valiente 

Department of Information and Communications Engineering, University of Murcia, Murcia, Spain

Correspondence: Mariano Albaladejo-González (mariano.albaladejog@um.es)

Received: 20 March 2024 | **Revised:** 24 October 2024 | **Accepted:** 22 January 2025

Funding: This work was supported by the Fundación Séneca, Grant/Award Number: 21795/FPI/22, 21948/JLI/22 and 22238/PDC/23; Instituto Nacional de Ciberseguridad, Grant/Award Number: CDL-TALENTUM.

Keywords: artificial intelligence | just-in-time interventions | learning analytics | machine learning | xAI

ABSTRACT

In recent years, serious games (SGs) have emerged as a powerful tool in education by combining pedagogy and entertainment, facilitating the acquisition of knowledge and skills in engaging environments. SGs enable the collection of valuable interaction data from students, allowing for the analysis of student performance, with artificial intelligence (AI) playing a key role in processing this data to make informed inferences about their knowledge and skills. However, the lack of explainability in AI models represents a significant challenge. This research aims to develop an interpretable model for predicting students' performance in real-time while playing an SG by: (1) calculating the performance of an interpretable prediction model of task completion in an SG and (2) demonstrating the application of the interpretable model for just-in-time (JIT) classroom interventions. Our results show that we are able to predict students' task completion in real-time with a balanced accuracy result of 77.21% after a short play-time has elapsed. In addition, an explainable artificial intelligence (XAI) approach has been applied to ensure the interpretability of the developed models. This approach supports personalised learning experiences, unlocks AI benefits for non-technical users, and maintains transparency in education.

1 | Introduction

In recent years, the use of games in educational contexts has significantly increased. Digital games provide an additional way for students to develop cognitive, spatial, and motor skills; help improve information and communication technology knowledge; teach complex problem-solving; and increase creativity, all while addressing topics that might be perceived as too complicated in a traditional classroom setting (Papanastasiou et al. 2017). Specifically, the potential of serious games (SGs)—games that do not have entertainment, enjoyment, or fun as their main purpose (Laamarti, Eid, and El Saddik 2014)—is particularly relevant, as they offer a unique blend of entertainment and

pedagogy. SGs can provide an excellent context that not only facilitates the acquisition and assessment of knowledge and skills but also allows for a detailed examination of environments free from the usual constraints of time and space (Bellotti, Berta, and De Gloria 2010). In the field of education, there is significant enthusiasm surrounding game-based assessment (GBA) due to the apparent limitations of conventional assessment methods in capturing students' knowledge, skills, and attributes (de Klerk and Kato 2017).

One of the key benefits of utilising SGs in educational settings is the wealth of data that can be collected from these interactive experiences, providing a great opportunity to make inferences

and assessments in ways that are not possible in traditional testing (Gomez, Ruipérez-Valiente, and Clemente 2023). The scope of the collected data can range from measuring individual skills at a granular level to evaluating attitudes at a larger scale. Moreover, the collected data can be intentionally designed to assess various aspects, such as knowledge, attitudes, skills, or behaviour (Smith, Blackmore, and Nesbitt 2015). Data from SGs can also be used by teachers to provide targeted assistance and support to students precisely when they need it. Specifically, just-in-time (JIT) instruction occurs when information, skill demonstration, or other necessary instruction is delivered on the spot at the time it is required, ensuring that information is available for immediate application in the relevant context (Anderson and Wood 2009).

The use of SGs in education does not necessarily imply teacher disengagement from teaching. Like any other learning situation, students require the guidance of their teachers during gameplay. It is the teacher's responsibility to ensure that all students are progressing through the game and successfully achieving both the game goals and the learning objectives (Bado 2022). However, monitoring students in real-time poses challenges that encompass several key aspects educators face in educational environments, such as conducting meaningful data analysis and identifying when and how to intervene based on the collected data. The use of artificial intelligence (AI) can support teachers in this task. In particular, monitoring students' performance in SGs generates a substantial amount of information that needs to be processed, and Machine Learning (ML) models can handle this data and identify patterns that allow us to recognise behaviours (Marín-Morales et al. 2021). That being said, the majority of existing research focuses on post hoc analyses or predicting overall student performance after finishing the game. Few studies have attempted to provide real-time insights that could inform JIT interventions.

Although ML models have demonstrated strong predictive capabilities in educational contexts, explainability remains an inherent problem of the latest techniques (e.g., ensembles or Deep Neural Networks) (Arrieta et al. 2020), and the increasing utilisation of ML models has led to a growing demand for transparency in AI (Preece et al. 2018). In this sense, explainable AI (XAI) proposes creating a suite of techniques that produce more explainable models while maintaining high performance. This enables end users to understand, appropriately trust, and effectively manage the emerging generation of AI models (Gunning 2017). In this way, XAI tools empower non-technical users (such as teachers) to interpret AI-generated insights and recommendations, enabling them to make informed decisions. Yet, there is a clear gap in research focusing on the combination of real-time predictive models and XAI in SGs. To the best of our knowledge, no previous studies have explored how interpretable models can be used to deliver JIT interventions during gameplay by predicting students' performance.

In this research, we aim to address these gaps by building an explainable model capable of predicting students' level outcomes in real-time while playing an SG. With this purpose in mind, we design a set of features derived from students' interaction with *Shadowspect*, a game designed as a formative assessment tool to assess mathematical content standards. Using these features,

we build a set of ML models that aim to predict the students' performance in real-time while ensuring that the models remain explainable. Unlike previous studies that have mainly focused on post-game performance predictions, our work makes early predictions during gameplay. In addition, we make results explainable and immediately applicable to JIT interventions. This real-time, interpretable feedback is a key contribution, as it enables educators to intervene promptly based on students' progress. Specifically, we have the following objectives:

1. **Configure and evaluate interpretable models of task completion at different time windows in an SG.** We aim to assess the model's performance by measuring its predictive accuracy and evaluating its ability to anticipate students' progress within the game environment. Furthermore, we will address the interpretability of the AI models to ensure that end users can interpret the results.
2. **Demonstrate the application of the interpretable model for JIT classroom interventions.** Our final objective is to showcase the practical application of the interpretable prediction model for real-time classroom interventions. We will conduct a use case that illustrates how our model can be applied in a real scenario.

The rest of the paper is structured as follows: Section 2 reviews background literature on SGs and assessment, AI models and XAI. Section 3 describes the methodology followed to conduct the research, as well as the game and the data collection used. Next, Section 4 presents the results, including the models developed, their interpretability analysis, and finally the case study conducted. Then, we finalise the paper with a discussion in Section 5 and conclusions and future work in Section 6.

2 | Related Work

The concept of games dates back to ancient civilisations and is recognised as a fundamental aspect of human societies throughout history (Laamarti, Eid, and El Saddik 2014). In particular, the versatility and adaptability of SGs make them a valuable tool across various contexts and domains. First of all, SGs have gained significant popularity in educational settings. For example, Ruipérez-Valiente et al. (2020) used "The Radix Endeavor" (an inquiry-based online game for STEM learning) in K-12 classrooms as part of a pilot study conducted in numerous schools. Additionally, SGs have also been proposed as a potential method for employee selection by improving the user experience, and the use of games in the workplace is a growing phenomenon, with SGs being increasingly used as evaluative tools (Al Qallawi and Raghavan 2022). Larson (2020) conducted a literature review on SGs and gamification in corporate environments, finding that the use of SGs is becoming increasingly prevalent. Moreover, SGs have been considered positive and innovative solutions for addressing contemporary issues in organisations, including meeting the needs of modern learners within the corporate context. In healthcare, SGs, particularly adventure and shooter games, play an important role in education, prevention and rehabilitation (Wiemeyer and Kliem 2012). In this regard, one of the most challenging areas is modelling simulations for

medical training. ‘CancerSpace’, developed by the National Cancer Institute and Oak Ridge Institute for Science and Education, is an SG that aims to facilitate cancer screening and consequently increases cancer-screening rates in federally qualified health centres (Swarz et al. 2010).

Moreover, predicting students’ performance in educational environments has gained considerable attention due to its potential to transform educational practices and enhance learning outcomes. Rastrollo-Guerrero, Gómez-Pulido, and Durán-Domínguez (2020) conducted a literature review of 70 papers to examine techniques and objectives for predicting students’ performance, noting a strong tendency to focus on university-level predictions (around 70% of the analysed articles). However, the authors also highlighted the need to apply these predictions at the school level, which would help identify low-performing students at earlier ages. Regarding SGs, there have been several studies attempting to analyse students’ data to predict their performance. For example, research presented by Illanas Vila et al. (2013) aimed to predict students’ performance in translating foreign languages by collecting data from 55 students in an SG and building neural network models. Although the results obtained were positive, the authors acknowledged certain limitations, such as the potential bias in their data set. Additionally, Kickmeier-Rust (2018) conducted a simulated study based on existing datasets using a multidimensional domain and learner models to add information about the nature of a learning domain. Furthermore, Abeyrathna et al. (2019) built a multi-label classifier using in-game data and player information to predict student proficiency in a quantum cryptography SG. More recent work by Hooshyar et al. (2023) attempted to predict early student performance using only 50% of learners’ action sequences, achieving a relative error of less than 8%. Finally, other studies have used ML models to analyse students’ interactions and predict their performance using in-game data (Yuhana et al. 2017; Loh, Sheng, and Li 2015; Lee et al. 2023; Alonso-Fernández et al. 2020). However, while these studies focused on performance prediction, none have demonstrated how to make their results explainable or immediately applicable to JIT interventions.

There are many institutions already using AI technologies to shape and plan the delivery of education (Zawacki-Richter et al. 2019). The use of AI has become a focal point for innovation and competitive advantage, with applications anticipated in areas such as learner profiling, intelligent tutoring systems, assessment, and personalised learning (Farrow 2023). XAI emerges to address the ‘black box’ perception non-technical users often have about AI, which can seem ‘humanly inexplicable’. As AI becomes more prevalent in education, XAI should help educators and learners understand the algorithms that influence the learning process. Previous research has examined the use of XAI in educational contexts. For example, Tao et al. (2020) presented an explainable multi-view game cheating detection framework driven by XAI.

Regarding the use of XAI for predicting student performance, Chitti, Chitti, and Jayabalan (2020) noted that the prediction models generated are often complex and not interpretable, making it difficult to understand why and how predictions are made based on the results. Thus, model interpretability has become

increasingly important. Alamri and Alharbi (2021) conducted a literature review investigating explainable models for predicting student performance from 2015 to 2020. Their findings revealed that the predictors used to train these models primarily consisted of a combination of socio-economic features and pre-course performance features. However, the review also highlighted that the potential of utilising e-learning analytics data as a source for explainable student performance models has not been fully explored. Nevertheless, we found some studies that used XAI to create interpretable student performance models. For example, Jang et al. (2022) applied several ML models to predict performance and verify whether at-risk students could be identified using selected features, providing helpful information to each student through XAI techniques. Regarding SGs, Berger and Müller (2021) designed a rule-based, short-term decision-making algorithm that reports game progress, demonstrating its suitability for creating adaptive SGs.

Table 1 provides a detailed comparison of previous studies on student performance prediction and XAI applications in SGs. Our research stands out from existing literature on predicting student performance in SGs by addressing two crucial aspects: real-time prediction and model explainability. Although Berger and Müller (2021) also tackled model interpretability and real-time prediction, their approach was entirely different, as they developed a rule-based model to predict in-game progress for adaptive SGs. In contrast, our work focuses on predicting student performance using ML models, with an emphasis on interpreting those predictions in real-time.

As observed, previous studies have attempted to forecast student performance in games, but these predictions have typically been made after the completion of the activity or game. Our research aims to push the boundaries by making real-time predictions using in-game data. By incorporating XAI techniques, we aim to shed light on the underlying factors and decision-making processes influencing student performance predictions.

3 | Methodology

Next, we present the SG *Shadowspect* along with our data collection in Section 3.1, followed by a detailed description of the complete process used to conduct the study in Section 3.2.

3.1 | Context and Dataset

In our research, we utilised *Shadowspect*, a 3D geometry game specifically designed to evaluate math core standards, including the visualisation of relationships between 2D and 3D objects. This allows teachers to integrate it into their core math curriculum. The game enables students to create composite figures using primitive shapes (e.g., pyramids, cones) and silhouettes from different perspectives. Players can manipulate shapes, change perspectives, and receive feedback on how well they match the silhouettes. An example of a puzzle being solved is shown in Figure 1. The current version of the game consists of 30 levels, divided into nine tutorial, nine intermediate and 12 advanced levels. While the tutorial levels focus on teaching

TABLE 1 | Detailed comparison of existing studies on student performance prediction and XAI applications in SGs.

| Study | Prediction goal | Methods | Data source | Dataset users | Interpretable model | Real-time prediction |
|--------------------------------|---|--|---------------------------------|---------------|---------------------|----------------------|
| Illanas Vila et al. (2013) | Predicting student performance in foreign languages | Neural networks | In-game data | 55 | X | X |
| Kickmeier-Rust (2018) | Predicting student performance in game-based scenarios | Linear regression model | Math competencies | 912 | ✓ | X |
| Abeyrathna et al. (2019) | Predicting student proficiency in quantum cryptography | Support vector machine | In-game and player data | 150 | X | X |
| Loh, Sheng, and Li (2015) | Predicting expert-novice performance differences | Partial least squares discriminant analysis | In-game data | 62 | X | X |
| Lee et al. (2023) | Predicting student posttest math knowledge scores | Seven ML models | In-game data | 359 | X | X |
| Yuhana et al. (2017) | Predicting student performance in math skills | Five ML models, one rule-based classifier | In-game answers and player data | 160 | X | X |
| Hooshyar et al. (2023) | Predicting student performance at early stages | Eight ML models, one deep learning (DL) model | In-game features | 427 | X | ✓ |
| Alonso-Fernández et al. (2020) | Predicting student knowledge given as post-test results | Decision trees, naïve bayes, logistic regression | In-game data | 227 | ✓ | X |
| Berger and Müller (2021) | Predicting in-game progress for adaptive SGs | Rule-based algorithm | In-game data | 80 | ✓ | ✓ |
| Our work | Predicting student-level performance in real-time | Seven ML models | In-game data | 322 | ✓ | ✓ |

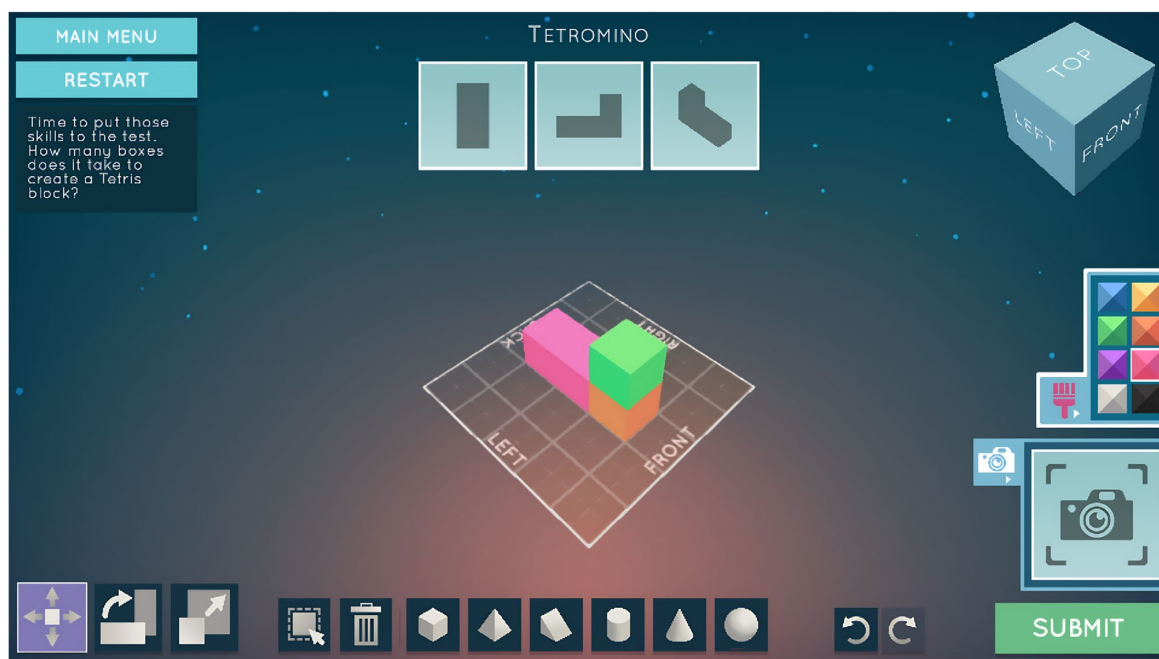


FIGURE 1 | Puzzle example in *Shadowspect*.

basic functionality, the intermediate and advanced levels offer more independence and challenging puzzles for experienced players.

For the data collection process, the team recruited seven teachers who used the games with students from seventh to tenth grade. The final dataset used for this research consists of approximately 428,000 events performed by a total of 322 students (with an average of 1320 events per student). These events were recorded over a period of 260h, equivalent to an average of 0.82h per student.

3.2 | Training and Explicability Procedure

We can divide the methodology into two main blocks: performance prediction model development and model interpretability and explanations. The first block describes the methodology used to build the ML models for performance prediction and consists of two stages: feature engineering and model training. In the feature engineering stage, a set of features was developed to assist in prediction, while the model training stage details how the ML models were trained. The second block focuses on the methodology used to enhance model interpretability, with the approach adapted according to the model's inherent interpretability. We can see a diagram illustrating the complete methodology in Figure 2.

3.2.1 | Feature Engineering

The first step was to design the features used for predicting users' performance in *Shadowspect*, specifically whether a user would successfully complete the level being played. To achieve this, we

employed a multi-prediction approach based on distinct time intervals derived from the average time required to complete each in-game puzzle. These intervals were set at 25%, 50%, and 75% of the average completion time, allowing us to monitor player progress and provide timely feedback or interventions as needed.

We designed a set of features crucial for predicting user success, categorised into three groups: **user features**, **puzzle features**, and **attempt features**. **User features** provide insights into the user's overall performance and interaction patterns, helping us understand their ability and strategies in the game. **Puzzle features** cover a set of data from the different levels within the game, allowing us to capture the unique characteristics and difficulty of each level, which is essential for accurate predictions. Finally, **attempt features** provide specific information about what the user is doing in that specific attempt, giving us a detailed view of the actions and decisions during gameplay. Using this comprehensive feature design allows us to consider both user general behaviour and the specifics of each gaming scenario. Table 2 presents the user features, Table 3 displays the puzzle features, and finally, Table 4 shows the attempt features. These features aim to summarise all aspects of users' interactions with *Shadowspect* during puzzle-solving attempts. It is important to note that **puzzle features** and **user features** do not vary with different time intervals since they aim to provide an overview of the user and the puzzle being played. Therefore, the features that vary with time intervals will be **attempt features**, as they provide information about the current game session.

Regarding the `user_elo` and `puzzle_elo` features, they are both calculated using an adapted version of the original Elo algorithm, which was initially designed as a method to rank chess players (Elo 2008). In our context, we consider the student and the puzzle as the opponents in our game, and each student's

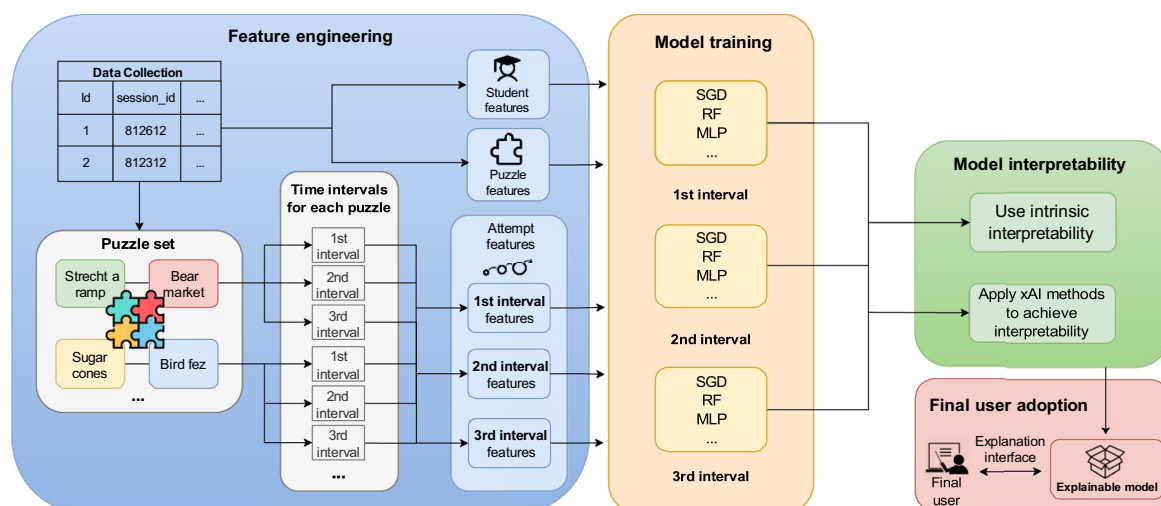


FIGURE 2 | Methodology diagram.

TABLE 2 | User features.

| Feature | Description |
|-------------------------|---|
| percentage_tutorial | Percentage of completed puzzles in the tutorial category. |
| percentage_intermediate | Percentage of completed puzzles in the intermediate category. |
| percentage_advanced | Percentage of completed puzzles in the advanced category. |
| attempts_per_puzzle | The average number of solution attempts needed by the user to solve a puzzle. |
| user_elo | Elo ranking obtained by facing each user against different puzzles. |

attempt at solving a puzzle is considered a match. A comprehensive description of how the adapted algorithm works can be checked in previous work (Ruipérez-Valient et al. 2022).

3.2.2 | Model Training

For training our models, we have considered the following algorithms:

- **Adaboost:** An ensemble method that combines weak learners to create a strong learner by iteratively adjusting weights.
- **Decision Tree (DT):** Creates a tree-like model of decisions by recursively splitting data based on different feature conditions.

TABLE 3 | Puzzle features.

| Feature | Description |
|-------------------|---|
| puzzle_difficulty | Calculated using the average time it takes to complete the puzzle, the average number of actions to solve it, the percentage of abandonments, and the percentage of incorrect checks when attempting to solve it are normalised separately with respect to the same metrics obtained from the rest of the puzzles. Once normalised, they are aggregated and normalised again with respect to the distribution of puzzles. |
| puzzle_elo | Elo ranking obtained by facing each puzzle against different users. |
| puzzle | The string containing the name of the puzzle. |

- **K-Nearest Neighbours (KNN):** Assigns a label to a new data point based on the labels of its k nearest neighbours.
- **Multi-Layer Perceptron (MLP):** A type of feedforward neural network with multiple layers to model complex non-linear relationships in data.
- **Random Forest (RF):** Ensemble method combining multiple decision trees to provide reliable predictions.
- **Stochastic Gradient Descent (SGD):** Efficient optimisation algorithm using small random subsets of training data, suitable for large datasets and online learning.

TABLE 4 | Attempt features.

| Feature | Description |
|-----------------------|--|
| n_events | Total number of events generated by the user. |
| n_breaks | Number of user's idle moments (15s without generating any events). |
| n_snapshot | Number of screenshots taken by the user. |
| n_rotate_view | Number of camera rotations performed by the user. |
| n_manipulation_events | Number of manipulation events generated by the user. |
| n_check_solution | Number of attempts to check the solution. |
| best_submit | Best puzzle submission rate obtained, determined by dividing the number of matched silhouettes by the total number of correct silhouettes. |

- **Support Vector Classifier (SVC):** Constructs hyperplanes to separate data into different classes, effective in high-dimensional spaces and for complex decision boundaries.

Regarding the data preprocessing, it is worth mentioning the separation of the dataset into training and testing sets. In this work, we needed to consider an additional condition beyond simply separating the data. We wanted to ensure that attempts from the same user fell into the same dataset to prevent the model from learning based on a user's specific behaviour and then predicting other attempts from the same user. This guarantees that the model is capable of generalising well to user behaviour across different individuals. Therefore, we decided to randomly select 70% of the users as training users and the remaining 30% as testing users. The attempts corresponding to the training users were included in the training dataset, while the attempts made by the testing users were included in the testing dataset. In addition, we applied one-hot encoding to the puzzle feature because some of the ML models used are unable to handle categorical data.

For model training and configuration, we employed ten-fold cross-validation and used balanced accuracy as the performance metric. This metric calculates the percentage of correctly classified positive and negative instances and then averages these percentages. It is highly useful in scenarios with imbalanced data, as it assigns equal importance to the accuracy of both the majority and minority classes. In our case, we are dealing with imbalanced data, as the number of instances corresponding to successes is nearly twice that of failures; thus, we prioritise balanced accuracy as our primary evaluation metric. In addition to the balanced accuracy, we also reported the F1 score, Matthews

correlation coefficient (MCC), precision, sensitivity and specificity. After training each model configuration separately, we selected the models and configurations associated with the algorithms that achieved the best average balanced accuracy for each time interval.

3.2.3 | Model Interpretability and Explanations

Once the best model was selected, we aimed to enhance its interpretability, ensuring that the model's predictions were fully explainable. If the chosen model was inherently interpretable, we would utilise that interpretability to explain the model's predictions. For example, if the best model happened to be a *decision tree*, we could easily interpret its decision-making process by examining the sequence of split rules and feature importance. On the other hand, if the chosen model was not interpretable, we sought to apply XAI methods to achieve interpretability.

XAI methods can be categorised based on different criteria. The first criterion is 'intrinsic vs. post hoc', where interpretability is achieved either by constraining the complexity of the ML model (intrinsic) or by using methods that analyse the model after training (post hoc). The second criterion is 'model-specific vs. model-agnostic'. Model-specific interpretation tools are limited to certain model classes, like interpreting regression weights in linear models. Model-agnostic tools, on the other hand, can be used with any ML model and are applied after the model has been trained. Lastly, the third criterion is 'local vs. global' interpretation. Local methods explain individual predictions, while global methods provide explanations for the overall behaviour of the entire model.

In this particular scenario, our intention was to use a post hoc, model-agnostic, and local interpretability method. We specifically opted for a post hoc approach because the chosen model was not inherently explainable. Furthermore, we aimed for it to be model-agnostic, as using a method tied to a particular model might not be compatible with the best-performing model. Lastly, we chose a local method, as our primary focus was on explaining individual predictions. The selected method for adding interpretability to a 'black box' model was *SHAP*. The primary reason for choosing *SHAP* was its capability to provide local interpretability. Additionally, *SHAP* values can be aggregated globally, offering insights into the model's overall behaviour and functioning.

SHAP is a method for explaining individual predictions based on Shapley values, which are optimal from the perspective of game theory. Shapley's values are a measure used in cooperative game theory to fairly allocate the value or contribution of each player to a specific game or problem. Shapley values are based on the idea that a player's value in a game depends on their contribution relative to the different possible coalitions or combinations of players. In other words, a player's value is calculated by considering all the possible ways they could have collaborated with the other players. To calculate Shapley's values, all possible permutations of players are considered, and the change in the game value when an additional player is added to the coalition is evaluated. These changes are averaged to obtain a fair measure of each player's value (Molnar 2022). In our context, the players are the different

features of the model, and the game value is the prediction outcome (the estimated success percentage of the attempt).

To calculate *SHAP* values, we selected the Python library *SHAP* (Lundberg 2018). Specifically, we employed the *explainer dashboard*, which builds upon *SHAP* to offer the option of visually representing *SHAP* values through dashboards. It is important to note that the importance of features in an individual prediction has a local interpretation. In other words, this importance is also influenced by the values of the remaining features. Therefore, in two different attempts where one feature holds the same value but the other features vary, they may have different levels of importance for the prediction. Thus, it is crucial not to attempt to explain the importance of a feature in isolation without considering the values of the other features, as these values also influence the significance of that particular feature.

4 | Results

4.1 | Performance Prediction Models: Comparison and Identification of the Best Model for Prediction

The first step after preprocessing was the configuration and evaluation of the AI algorithms through a ten-fold cross-validation for each time interval. The best results obtained by each algorithm and time interval in the cross-validations are summarised in Tables 5–7. The results show that RF achieved the best performance in every time interval, achieving a balanced accuracy of 0.76 in the first time interval (25%), 0.772 in the second time interval (50%), and 0.795 in the third time interval (75%). It is noteworthy that the algorithms achieved high performance even in the first time interval, and the performance improvement when moving to the second and third

TABLE 5 | Ten-fold cross-validation training results for prediction 1 (25th).

| Model | Balanced accuracy | F1 score | MCC | Precision | Sensitivity | Specificity |
|----------|-------------------|----------|--------|-----------|-------------|-------------|
| AdaBoost | 0.7253 | 0.8887 | 0.5069 | 0.8513 | 0.9324 | 0.5183 |
| DT | 0.7448 | 0.8857 | 0.5106 | 0.9103 | 0.5794 | 0.8637 |
| KNN | 0.6746 | 0.8731 | 0.4092 | 0.8253 | 0.9297 | 0.4195 |
| MLP | 0.7408 | 0.8858 | 0.5077 | 0.8628 | 0.9118 | 0.5697 |
| RF | 0.7598 | 0.8801 | 0.5175 | 0.8824 | 0.8793 | 0.6403 |
| SGD | 0.7230 | 0.7548 | 0.3897 | 0.9029 | 0.6611 | 0.7848 |
| SVC | 0.6964 | 0.8314 | 0.3690 | 0.8551 | 0.8112 | 0.5816 |

TABLE 6 | Ten-fold cross-validation training results for prediction 2 (50th).

| Model | Balanced accuracy | F1 score | MCC | Precision | Sensitivity | Specificity |
|----------|-------------------|----------|--------|-----------|-------------|-------------|
| AdaBoost | 0.7621 | 0.8782 | 0.5722 | 0.8337 | 0.9309 | 0.5932 |
| DT | 0.7433 | 0.8293 | 0.4874 | 0.8187 | 0.6680 | 0.8460 |
| KNN | 0.6995 | 0.8371 | 0.4311 | 0.7949 | 0.8866 | 0.5124 |
| MLP | 0.7653 | 0.8620 | 0.5493 | 0.8430 | 0.8853 | 0.6453 |
| RF | 0.7715 | 0.8698 | 0.5635 | 0.8526 | 0.8911 | 0.6518 |
| SGD | 0.7096 | 0.7854 | 0.4118 | 0.8340 | 0.7529 | 0.6663 |
| SVC | 0.7183 | 0.8557 | 0.4891 | 0.8129 | 0.9068 | 0.5299 |

TABLE 7 | Ten-fold cross-validation training results for prediction 3 (75th).

| Model | Balanced accuracy | F1 score | MCC | Precision | Sensitivity | Specificity |
|----------|-------------------|----------|--------|-----------|-------------|-------------|
| AdaBoost | 0.7789 | 0.8261 | 0.5659 | 0.7908 | 0.8716 | 0.6863 |
| DT | 0.7565 | 0.7787 | 0.4981 | 0.7749 | 0.7380 | 0.7873 |
| KNN | 0.7217 | 0.7897 | 0.4604 | 0.7300 | 0.8676 | 0.5759 |
| MLP | 0.7885 | 0.8401 | 0.5985 | 0.7980 | 0.8921 | 0.6849 |
| RF | 0.7954 | 0.8383 | 0.6001 | 0.8090 | 0.8737 | 0.7172 |
| SGD | 0.7035 | 0.7271 | 0.4164 | 0.7394 | 0.7663 | 0.6407 |
| SVC | 0.7365 | 0.8087 | 0.5009 | 0.7488 | 0.8840 | 0.5889 |

time intervals is not particularly high. The main reasons are that the user and puzzle features contain relevant information for predicting student performance, and the attempt features quickly capture the student's skills.

Once we selected RF as the best algorithm, we assessed the generalisation power of the three RF models (one for each time interval) in the test set. The test set contained unseen data not employed during the models' training, configuration and selection. Table 8 shows the balanced accuracy achieved by RF in the test set of each time interval. The performance of RF was even better than in the cross-validation, showing a good generalisation power of the three models.

Based on the results mentioned above, we chose RF to be applied in the use case that we will present afterward. However, RF is a 'black box' algorithm. Therefore, our models are not inherently interpretable, unlike other algorithms such as KNN or decision trees. To improve the interpretability of the models, we applied the SHAP method, which allows us to explain individual predictions in the use case. Moreover, this method can also be useful for identifying the most relevant factors contributing to students' success when solving puzzles. To achieve this, we rely on the importance of features calculated as the average absolute value of the Shapley values obtained by each feature in different individual predictions. We can see the importance of each feature in each prediction in Figure 3. **Puzzle features** are shown in red, **user**

TABLE 8 | Test results.

| Model | Balanced accuracy | F1 score | MCC | Precision | Sensitivity | Specificity |
|---------------------|-------------------|----------|--------|-----------|-------------|-------------|
| Prediction 1 (25th) | 0.7721 | 0.8445 | 0.5439 | 0.8449 | 0.8441 | 0.7 |
| Prediction 2 (50th) | 0.7918 | 0.8288 | 0.5912 | 0.8026 | 0.8567 | 0.7268 |
| Prediction 3 (75th) | 0.7928 | 0.8020 | 0.5877 | 0.7723 | 0.8340 | 0.7516 |

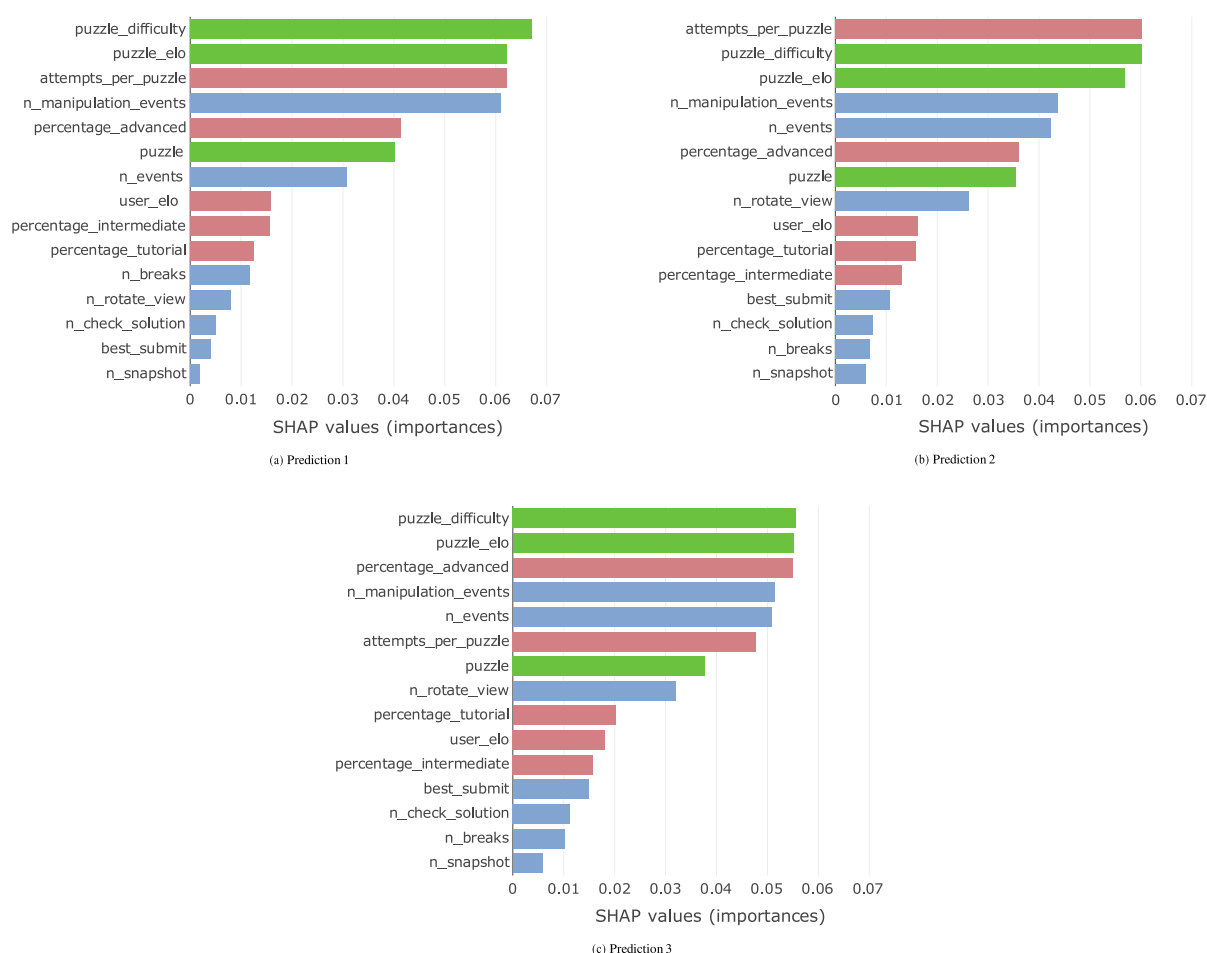


FIGURE 3 | Importance of each feature for each prediction.

features in green and **attempt features** in blue. Although there are a few exceptions, the relevance of most features remains relatively stable across predictions.

We observe that the most crucial features across all predictions are puzzle-related, specifically `puzzle_difficulty` and `puzzle_elo`. This implies that the decisive factor in determining whether a user will solve a puzzle is the puzzle's difficulty. Moving to **user features**, both the percentage of advanced puzzles completed (`percentage_advanced`) and the number of attempts per puzzle (`attempts_per_puzzle`) hold significant importance. In contrast, the percentage of intermediate puzzles completed (`percentage_intermediate`) and the user Elo (`user_elo`) show low relevance. This suggests that efficient students who have successfully tackled advanced puzzles and made few mistakes are more likely to solve the puzzle they currently face. Finally, regarding **attempt features**, we note that the number of events (`n_events` and `n_manipulation_events`) plays a crucial role in prediction. Therefore, in order to successfully solve a puzzle, it is essential that the user is proactive, actively creating and manipulating figures.

Finally, we can take a look at the remaining **attempt features**: `n_check_solution`, `n_rotate_view`, `n_breaks`, `best_submit` and `n_snapshot`. All these features have low relevance in the prediction, except for the number of camera rotations, which gains relevance in the third prediction. Considering this, it is not crucial whether a student takes prolonged pauses during puzzle-solving, the number of screenshots taken, or the best solution achieved up to that moment. However, we can interpret that performing more camera rotations as the solving process advances helps users find a solution when stuck. This action offers new perspectives on the scenario, which is a crucial aspect in spatially related geometric problems.

4.2 | Use Case: Supporting Individual Students in the Classroom

For this case study, we envision a classroom scenario in which a teacher is using the tool and has access to a dashboard to monitor struggling students. This dashboard can identify students with low probabilities of completing a level, enabling the teacher to prioritise them for JIT support and assistance in completing the task. To do so, the teacher can analyse the individual predictions of the model displayed on the dashboard to better understand its functioning. We will present two examples from our dataset, one where the model corresponding to the second prediction correctly predicts a failed attempt and another where it correctly predicts the success of an attempt. Additionally, we will analyse the obtained *SHAP* values for each feature to study how the teacher can interpret these values and assist students appropriately.

Regarding the first student, Table 9 shows the features associated with this student's second attempt at this level, along with the obtained *SHAP* values for each feature. The user who made this attempt has completed 88.89% of the tutorial puzzles (`percentage_tutorial`), none of the intermediate

TABLE 9 | Features and *SHAP* values of the first student's attempt.

| Feature | Value | SHAP |
|--------------------------------------|-------------|---------|
| <code>percentage_tutorial</code> | 88.89 | −0.0161 |
| <code>percentage_intermediate</code> | 0.00 | 0.0157 |
| <code>percentage_advanced</code> | 7.69 | 0.0019 |
| <code>attempts_per_puzzle</code> | 2.33 | −0.0648 |
| <code>user_elo</code> | 0.08 | −0.0014 |
| <code>puzzle_elo</code> | 3.00 | −0.126 |
| <code>puzzle_difficulty</code> | 1.00 | −0.1265 |
| <code>n_events</code> | 17.00 | 0.0176 |
| <code>n_check_solution</code> | 0.00 | 0.0056 |
| <code>best_submit</code> | 0.00 | −0.0064 |
| <code>n_breaks</code> | 0.00 | −0.0008 |
| <code>n_manipulation_events</code> | 6.00 | 0.0228 |
| <code>n_snapshot</code> | 0.00 | −0.0034 |
| <code>n_rotate_view</code> | 0.00 | −0.0193 |
| <code>puzzle</code> | Bear market | −0.0777 |
| <code>completed</code> | 0.0 | |

puzzles(`percentage_intermediate`), and 7.89% of the advanced puzzles(`percentage_advanced`). The student has a `user_elo` value of 0.08 and an average of 2.33 attempts per completed puzzle (`attempts_per_puzzle`).

The puzzle that the student is trying to solve is 'Bear market', which is the most difficult puzzle in the entire game according to the `puzzle_elo` and `puzzle_difficulty` features. The user has performed 17 events (`n_events`), out of which six are manipulation events (`n_manipulation_events`). The value of `n_breaks`, `n_check_solution`, `n_rotate_view` and `n_snapshot` features is zero. This means that there have been no periods of inactivity, no attempts to check the puzzle solution, and no camera rotations or screenshots taken. The last row of the table indicates that the user did not manage to complete the puzzle in that particular attempt.

In Figure 4, the *SHAP* values and how they contribute to the final prediction are visually presented. The first bar indicates the model's prediction result without considering any features. As expected, this bar indicates 50% since, without any features, the model cannot determine whether to predict success or failure. Remember that if the final prediction percentage is higher than 50%, it will predict a successful resolution (one), and otherwise, it predicts failure (zero).

The following bars correspond to the obtained *SHAP* values (Table 9). Starting from the initial prediction percentage (50%), the corresponding *SHAP* values are added or subtracted until reaching the final prediction percentage (e.g., a *SHAP* value of 0.025 translates to a 2.5% increase in the final prediction). Green bars represent positive contributions, while red bars represent negative contributions. The most relevant

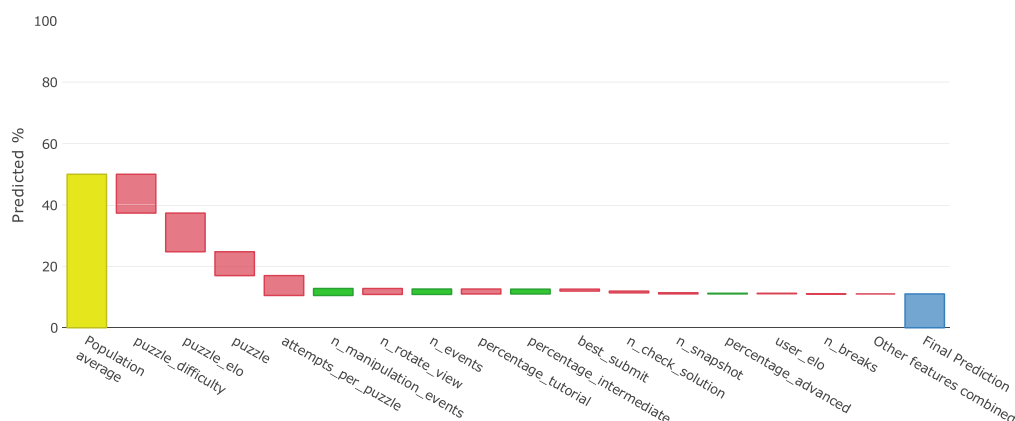


FIGURE 4 | SHAP values of the first student's attempt.

features in this prediction are related to the puzzle. Given that it is the most challenging puzzle in the game, the contribution of each of these features is highly negative to the final prediction. Additionally, the user has a high average number of attempts per completed puzzle (`attempts_per_puzzle`), which also has a significant negative impact on the prediction. Consequently, the model predicts a success percentage of only 11.03%, accurately predicting the user's failure to solve the puzzle.

Given this information provided by the dashboard, the teacher has the option to suggest that students start tackling easier puzzles since this student has begun solving the most difficult puzzle in the game. Thus, the teacher may advise them to concentrate on solving simpler puzzles first, such as the intermediate ones, especially if they have not completed any of them.

Now, we analyse an attempt from a different student. In Table 10, we can observe the features corresponding to the first attempt of this student in this particular puzzle. The user who played this attempt completed 55.6% of the tutorial puzzles and none of the intermediate or advanced puzzles. The student has a `user_elo` value of 0.0 (Elo does not consider tutorial puzzles) and an average of 1.0 attempt per completed puzzle, meaning that the user has successfully completed all the previous puzzles in the first attempt. Regarding the puzzle, in this case, it is '6. Stretch a ramp', which has a difficulty value of 0.11 and an Elo value of 0.42, making it relatively easy to solve. Up to that moment, the user has performed a total of 31 events (of which 15 were manipulation events), has had no periods of inactivity and neither rotated the camera nor took screenshots. Finally, the last row of the table indicates that the user successfully solved the puzzle in this attempt.

In Figure 5, the SHAP values and their contributions to the final prediction are visually presented. Starting from the initial prediction percentage (50%), we observe that the feature contributing the most is the number of attempts per completed puzzle. Since this user has completed all attempted puzzles on the first try, the model assigns significant relevance to this feature. Subsequently, puzzle-related features also contribute significantly to the decision, given the low difficulty of this puzzle. Finally, it is worth noting the relevance of manipulation events.

TABLE 10 | Features and SHAP values of the second student's attempt.

| Feature | Value | SHAP |
|-------------------------|------------------|---------|
| percentage_tutorial | 55.56 | 0.0205 |
| percentage_intermediate | 0.00 | 0.0163 |
| percentage_advanced | 0.00 | 0.03 |
| attempts_per_puzzle | 1.00 | 0.0893 |
| user_elo | 0.00 | 0.0189 |
| puzzle_elo | 0.42 | 0.0725 |
| puzzle_difficulty | 0.11 | 0.0713 |
| n_events | 31.00 | 0.0364 |
| n_check_solution | 0.00 | 0.0069 |
| best_submit | 0.00 | -0.0097 |
| n_breaks | 0.00 | 0.0073 |
| n_manipulation_events | 15.00 | 0.0497 |
| n_snapshot | 0.00 | -0.0003 |
| n_rotate_view | 0.00 | -0.0282 |
| puzzle | 6 Stretch a ramp | 0.0414 |
| completed | 1.0 | |

The user has extensively created and manipulated figures, a clear indication of being close to a puzzle solution, thus contributing positively to the model's prediction. In the end, the model predicts that the student will solve the puzzle with a probability of 92.3%. Given this scenario, the teacher does not need to assist the student due to the context provided by the features and the high probability of solving the puzzle.

5 | Discussion

The analysis and interpretation of data generated by SGs can provide valuable information for learners and instructors in educational settings. For example, instructors can follow a student's

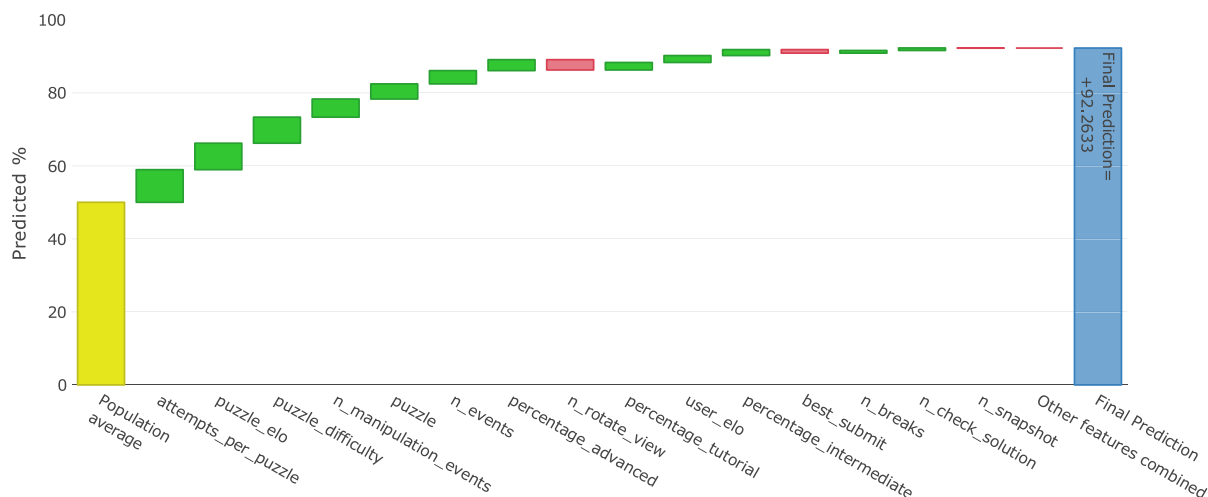


FIGURE 5 | SHAP values of the second student's attempt.

progression in real-time while playing and take action on any identified learning problems (Serrano-Laguna et al. 2017). This data analysis would enable the application of JIT interventions in classrooms. By continuously monitoring and interpreting students' interactions, game analytics allow instructors to identify learning gaps and challenges as they emerge. However, the analyses typically used in GBA studies are quite simple, and there is still a present challenge in developing more sophisticated and robust methods for leveraging the vast amount of data generated by SGs (Gomez, Ruipérez-Valiente, and Clemente 2023).

In this regard, AI models can be used to predict future performance based on current learning behaviours, adding a novel dimension to the personalisation and adaptation of GBAs (Kickmeier-Rust 2018). In this research, we developed a predictive approach that integrates AI as a crucial component in analysing this game data, taking into account several features derived from students' interactions with the game and allowing us to assess and predict students' performance while playing. These predictions can be useful in many ways. For instance, they enable instructors to intervene in the classroom when necessary, identifying students who are experiencing difficulties and helping them in real-time. Moreover, this information could also be used as input in adaptive platforms to adjust the game difficulty based on the predictions, provide hints and adjust the learning experience to each student's needs.

However, for instructors to provide this personalised learning approach, they need to understand the information generated by the AI models. XAI plays a crucial role in facilitating the socio-cultural process of learning, where interactions between teachers and students are fundamental in guiding learners through zones of proximal development and providing personalised support to students facing difficulties (Khosravi et al. 2022). Incorporating XAI methods further enhances the educational benefits of our proposal. XAI techniques facilitate the comprehension and explanation of how our AI models make predictions, and this transparency enables instructors to understand the reasoning the model followed to make the prediction and act accordingly.

Moreover, an important aspect of AI systems is to ensure user trust. Explainability gives users confidence that AI systems work well, helps developers understand why a system works a certain way, and safeguards against bias (Shin 2021). In education, the need for explanation arises since educators must be accountable to students, parents, and the government. Explanation is crucial when providing individual feedback to students, offering teachers diagnostic feedback to identify areas where a class of students needs increased focus, and during parental consultations to help them support their child's learning (Khosravi et al. 2022). By incorporating XAI techniques on top of our models, we increase the transparency of our predictive models in educational contexts, allowing instructors to understand the reasoning behind the predictions and building confidence in the potential use of SGs in the classroom.

6 | Conclusions

This research aimed to develop an XAI model for predicting students' performance in real-time while playing *Shadowspect*, a geometry SG. The RF predictive model developed in this study demonstrates a promising accuracy in anticipating students' task completion, achieving a balanced accuracy result of 77.21% in making early predictions after a short playtime has elapsed. Moreover, we ensured that the model predictions are fully explainable by taking into account both intrinsic and extrinsic explainability options. This way, our work provides a comprehensive framework for interpretable models, enabling a better understanding of the AI model predictions and facilitating informed decision-making in educational contexts. With this research, we aim to contribute to the educational field by providing a powerful and understandable tool that supports personalised learning experiences and effectively integrates SGs into educational settings.

This work has some limitations. First, our ML models were built using a specific set of log data from *Shadowspect*, which may not fully capture the diversity of behaviours present in other contexts and SGs. Expanding the feature set to include

data from other SGs could improve the models' applicability across different games and enhance predictive accuracy. Moreover, although our interpretable models are useful for understanding the model's decisions, the output might still be complex for non-technical users. Thus, we plan to refine the model's interpretability to ensure that even users with little technical expertise can benefit from the insights provided by the models. Furthermore, while we acknowledge the effectiveness of the ML models developed, the use of more complex techniques, such as DL, could potentially enhance the predictive power and robustness of our approach. Another limitation is that our data comes from a partially controlled classroom setting. As part of our future work, we intend to conduct case studies and experiments in various real-world educational settings to better understand the adaptability and generalisability of our approach. By addressing these limitations, we aim to make our model a valuable tool for personalised learning and interventions.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement


Research data are not shared.

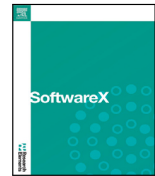
References

- Abeyrathna, D., S. Vadla, V. Bommanapally, M. Subramaniam, P. Chundi, and A. Parakh. 2019. "Analyzing and Predicting Player Performance in a Quantum Cryptography Serious Game." In *Games and Learning Alliance. GALA 2018. Lecture Notes in Computer Science*, edited by M. Gentile, M. Allegra, and H. Söbke, vol. 11385. Cham, Switzerland: Springer.
- Al Qallawi, S., and M. Raghavan. 2022. "A Review of Online Reactions to Game-Based Assessment Mobile Applications." *International Journal of Selection and Assessment* 30, no. 1: 14–26.
- Alamri, R., and B. Alharbi. 2021. "Explainable Student Performance Prediction Models: A Systematic Review." *IEEE Access* 9: 33132–33143.
- Alonso-Fernández, C., I. Martínez-Ortiz, R. Caballero, M. Freire, and B. Fernández-Manjón. 2020. "Predicting Students' Knowledge After Playing a Serious Game Based on Learning Analytics Data: A Case Study." *Journal of Computer Assisted Learning* 36, no. 3: 350–358.
- Anderson, A., and E. Wood. 2009. "Implementing Technology in the Classroom: Assessing Teachers' Needs Through the Use of a Just-In-Time Support System." In *Society for Information Technology & Teacher Education International Conference*, 3369–3372. Charleston, SC, USA: Association for the Advancement of Computing in Education (AACE).
- Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (Xai): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible Ai." *Information Fusion* 58: 82–115.
- Bado, N. 2022. "Game-Based Learning Pedagogy: A Review of the Literature." *Interactive Learning Environments* 30, no. 5: 936–948.
- Bellotti, F., R. Berta, and A. De Gloria. 2010. "Designing Effective Serious Games: Opportunities and Challenges for Research." *International Journal of Emerging Technologies in Learning (IJET)* 5, no. 2010: 22–35.
- Berger, F., and W. Müller. 2021. "Back to Basics: Explainable Ai for Adaptive Serious Games." In *Serious Games. JCSG 2021. Lecture Notes in Computer Science*, edited by B. Fletcher, M. Ma, S. Göbel, J. B. Hauge, and T. Marsh, vol. 12945. Cham, Switzerland: Springer.
- Chitti, M., P. Chitti, and M. Jayabalan. 2020. "Need for Interpretable Student Performance Prediction." In *2020 13th International Conference on Developments in Systems Engineering (Dese)*, 269–272. IEEE: Liverpool, United Kingdom.
- de Klerk, S., and P. M. Kato. 2017. "The Future Value of Serious Games for Assessment: Where Do We Go Now?" *Journal of Applied Testing Technology* 18, no. S1: 32–37.
- Elo, A. 2008. *The Rating of Chessplayers: Past and Present*. Bronx, NY: Ishi Press International.
- Farrow, R. 2023. "The Possibilities and Limits of Xai in Education: A Socio-Technical Perspective." *Learning, Media and Technology* 48, no. 2: 266–279.
- Gomez, M. J., J. A. Ruipérez-Valiente, and F. J. G. Clemente. 2023. "A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges." *IEEE Transactions on Learning Technologies* 16, no. 4: 500–515.
- Gunning, D. 2017. "Explainable Artificial Intelligence (XAI)." *Defense Advanced Research Projects Agency (DARPA), nd Web* 2, no. 2: 1.
- Hooshyar, D., N. El Mawas, M. Milrad, and Y. Yang. 2023. "Modeling Learners to Early Predict Their Performance in Educational Computer Games." *IEEE Access* 11: 20399–20417.
- Illanas Vila, A., J. Calvo Ferrer, F. Gallego Durán, and F. Llorens Largo. 2013. "Predicting Student Performance in Translating Foreign Languages With a Serious Game." In *INTED2013 Proceedings*, 52–59. IATED: Valencia, Spain.
- Jang, Y., S. Choi, H. Jung, and H. Kim. 2022. "Practical Early Prediction of Students' Performance Using Machine Learning and Explainable AI." *Education and Information Technologies* 27, no. 9: 12855–12889.
- Khosravi, H., S. B. Shum, G. Chen, et al. 2022. "Explainable Artificial Intelligence in Education." *Computers and Education: Artificial Intelligence* 3: 100074.
- Kickmeier-Rust, M. D. 2018. "Predicting Learning Performance in Serious Games." In *Serious Games: 4th Joint International Conference, JCSG 2018, Darmstadt, Germany, November 7–8, 2018, Proceedings*, vol. 4, 133–144. Cham: Springer International Publishing.
- Laamarti, F., M. Eid, and A. El Saddik. 2014. "An overview of serious games." *International Journal of Computer Games Technology* 2014: 1–15.
- Larson, K. 2020. "Serious Games and Gamification in the Corporate Training Environment: A Literature Review." *TechTrends* 64, no. 2: 319–328.
- Lee, J.-E., A. Jindal, S. N. Patki, A. Gurung, R. Norum, and E. Ottmar. 2023. "A Comparison of Machine Learning Algorithms for Predicting Student Performance in an Online Mathematics Game." *Interactive Learning Environments* 32, no. 9: 5302–5316.
- Loh, C. S., Y. Sheng, and I.-H. Li. 2015. "Predicting Expert–Novice Performance as Serious Games Analytics With Objective-Oriented and Navigational Action Sequences." *Computers in Human Behavior* 49: 147–155.
- Lundberg, S. 2018. "Shap Documentation." <https://shap.readthedocs.io/en/latest/index.html>.
- Marín-Morales, J., L. A. Carrasco-Ribelles, M. Alcañiz, and I. A. C. Giglioli. 2021. "Applying Machine Learning to a Virtual Serious Game for Neuropsychological Assessment." In *2021 IEEE Global Engineering Education Conference (Educon)*, 946–949. IEEE: Vienna, Austria.
- Molnar, C. 2022. *Interpretable Machine Learning*. 2nd ed. Victoria, Canada: Leanpub. <https://christophm.github.io/interpretable-ml-book>.
- Papanastasiou, G., A. Drigas, C. Skianis, and M. D. Lytras. 2017. "Serious Games in k-12 Education: Benefits and Impacts on Students With Attention, Memory and Developmental Disabilities." *Program* 51, no. 4: 424–440.

- Preece, A. D., D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty. 2018. "Stakeholders in Explainable AI." *CoRR*, abs/1810.00184.
- Rastrullo-Guerrero, J. L., J. A. Gómez-Pulido, and A. Durán-Domínguez. 2020. "Analyzing and Predicting students' Performance by Means of Machine Learning: A Review." *Applied Sciences* 10, no. 3: 1042.
- Ruipérez-Valient, J. A., Y. J. Kim, R. S. Baker, P. A. Martínez, and G. C. Lin. 2022. "The Affordances of Multivariate Elo-Based Learner Modeling in Game-Based Assessment." *IEEE Transactions on Learning Technologies* 16, no. 2: 152–165.
- Ruipérez-Valiente, J. A., M. Gaydos, L. Rosenheck, Y. J. Kim, and E. Klopfer. 2020. "Patterns of Engagement in an Educational Massively Multiplayer Online Game: A Multidimensional View." *IEEE Transactions on Learning Technologies* 13, no. 4: 648–661.
- Serrano-Laguna, Á., I. Martínez-Ortiz, J. Haag, D. Regan, A. Johnson, and B. Fernández-Manjón. 2017. "Applying Standards to Systematize Learning Analytics in Serious Games." *Computer Standards & Interfaces* 50: 116–123.
- Shin, D. 2021. "The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable Ai." *International Journal of Human-Computer Studies* 146: 102551.
- Smith, S. P., K. Blackmore, and K. Nesbitt. 2015. "A Meta-Analysis of Data Collection in Serious Games Research." In *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, 31–55. Cham: Springer International Publishing.
- Swarz, J., A. Ousley, A. Magro, et al. 2010. "Cancerspace: A Simulation-Based Game for Improving Cancer-Screening Rates." *IEEE Computer Graphics and Applications* 30, no. 1: 90–94. <https://doi.org/10.1109/MCG.2010.4>.
- Tao, J., Y. Xiong, S. Zhao, et al. 2020. "Xai-Driven Explainable Multi-View Game Cheating Detection." In *2020 IEEE Conference on Games (Cog)*, 144–151. IEEE: Osaka, Japan.
- Wiemeyer, J., and A. Kliem. 2012. "Serious Games in Prevention and Rehabilitation—A New Panacea for Elderly People?" *European Review of Aging and Physical Activity* 9, no. 1: 41–50.
- Yuhana, U. L., R. G. Mangowal, S. Rochimah, E. M. Yuniarno, and M. H. Purnomo. 2017. "Predicting Math Performance of Children With Special Needs Based on Serious Game." In *2017 IEEE 5th International Conference on Serious Games and Applications for Health (Segah)*, 1–5. Perth, WA, Australia: IEEE.
- Zawacki-Richter, O., V. I. Marín, M. Bond, and F. Gouverneur. 2019. "Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators?" *International Journal of Educational Technology in Higher Education* 16, no. 1: 1–27.

5 Optimizing Manual Labeling in GBA

| | |
|--|-----------------------------|
| Title Optimizing multimedia and gameplay data labeling: A web-based tool for Game-Based Assessment | |
| Authors <u>Manuel J. Gomez</u> ¹ , José A. Ruipérez-Valiente ¹ , Félix J. García Clemente ¹ ¹ <i>Department of Information and Communications Engineering, University of Murcia, Spain</i> | |
| Publication details | |
| Journal | SoftwareX |
| Volume | 27 |
| Pages | 101763 |
| JIF | 2.4 |
| Status | Published |
| Publisher | Elsevier |
| Number | – |
| Year | 2024 |
| Rank | Q2 |
| DOI | 10.1016/j.softx.2024.101763 |
| Abstract In this research, we introduce a novel open-source labeling tool, the Game-Based Assessment (GBA) Labeling Tool, specifically designed to address current challenges for data labeling in GBA scenarios. This web-based application facilitates the annotation of audio, video, and game event data, offering three different types of annotations – global, time instant, and time window annotations – to enhance accuracy in the labeling process. The tool also offers customizable labels and various types of visualizations to support different contexts and scenarios. | |
|  | |



Original software publication

Optimizing multimedia and gameplay data labeling: A web-based tool for Game-Based Assessment

Manuel J. Gomez ^{*}, José A. Ruipérez-Valiente, Félix J. García Clemente

Facultad de Informática, Universidad de Murcia, Murcia, Spain



ARTICLE INFO

Dataset link: <https://github.com/CyberDataLab/gba-labeling-tool/blob/main/sampleData>

Keywords:

Game-Based Assessment
Data annotation
Data labeling
Human labeling
Multimedia data

ABSTRACT

In this research, we introduce a novel open-source labeling tool, the Game-Based Assessment (GBA) Labeling Tool, specifically designed to address current challenges for data labeling in GBA scenarios. This web-based application facilitates the annotation of audio, video, and game event data, offering three different types of annotations – global, time instant, and time window annotations – to enhance accuracy in the labeling process. The tool also offers customizable labels and various types of visualizations to support different contexts and scenarios.

Code metadata

| | |
|---|---|
| Current code version | v1.0 |
| Permanent link to code/repository used for this code version | https://github.com/ElsevierSoftwareX/SOFTX-D-24-00172 |
| Permanent link to Reproducible Capsule | – |
| Legal Code License | MITLicense |
| Code versioning system used | git |
| Software code languages, tools, and services used | CSS, JavaScript, HTML, Python, Django |
| Compilation requirements, operating environments & dependencies | Ubuntu/MacOS, Python ≥ 3.7.2 |
| If available Link to developer documentation/manual | https://github.com/CyberDataLab/gba-labeling-tool/blob/main/readme.md |
| Support email for questions | manueljesus.gomez@um.es |

1. Motivation and significance

Modern technologies are having a significant impact on every industry, including gaming and education [1]. Serious Games (SGs), which are designed specifically for purposes other than or in addition to pure entertainment [2], have gained significant attention in recent years. In particular, SGs are being explored for their potential to provide more valid assessments compared to traditional assessment approaches. Game-Based Assessment (GBA) has been increasingly used in various domains such as education, health, military, and industry [3]. With the growing popularity and adoption of GBA, there has also been a significant increase in the quantity of data generated from user interaction.

Data collected for SGs and assessment research covers diverse measurements, including performance skills and behavioral factors relevant to both the process and outcomes [4]. Usually, various data types are collected, including audio recordings, video captures, and game event data from players' interaction with games. In this regard, labeled data plays a crucial role in the development of models and algorithms that enable researchers to gain deeper insights into learners' behaviors, engagement, and performance [5]. However, the process of data labeling is time-consuming and challenging. In many real-world scenarios, large-scale labeled datasets can be very costly to acquire, specially when expert annotators are required [6]. As datasets for training and testing of algorithms get increasingly larger, there is a need for efficient and user-friendly solutions to facilitate the annotation of multimedia data [7].

^{*} Corresponding author.

E-mail address: manueljesus.gomez@um.es (Manuel J. Gomez).

<https://doi.org/10.1016/j.softx.2024.101763>

Received 14 March 2024; Received in revised form 26 April 2024; Accepted 11 May 2024

Available online 20 May 2024

2352-7110/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

In the context of GBA, where large quantities of data are collected, the lack of standardized tools and methodologies for data labeling raises significant challenges. Researchers usually employ primitive methods such as Excel worksheets or manual annotations, and the existing multi-functional labeling tools are too general to be used with GBA data. Wang et al. [8] reported the very limited existence of labeled datasets in many real-world scenarios, and specially in education. They also noted that labels are often annotated by multiple workers with different expertise, leading to noise and inconsistency. Although we identified that many previous SGs studies have applied AI models/techniques [9,10], the majority of these studies heavily rely on automatically labeled data derived from in-game measures or statistics. Therefore, we note that human-labeled data is often underutilized, potentially due to the considerable time and economic resources required for its acquisition.

With this tool, we aim to address existing challenges associated with data annotation in GBA by developing a web-based open-source tool specifically designed for GBA data annotation. By integrating the possibility of labeling audio, video, and game-event data in the same tool, we provide researchers with a flexible and efficient solution for labeling various types of multimedia data. It provides different annotation types (global, time instant and time window annotations) and different ways of visualizing the same data. In addition, the tool's interface and workflow optimize the annotation process, potentially reducing inconsistencies between annotations and users' interpretations. By developing this tool, we contribute to the GBA literature and enhance the potential for leveraging multimedia data for educational research and assessment.

A comparative analysis with already existing tools in the literature reveals the unique features and enhancements of our tool. Zhang et al. [11] developed a conceptual framework for data labeling, and Da Silva et al. [7] presented an open source multipurpose tool for the annotation of multimedia datasets with collaborative annotation capabilities. While this multipurpose tool offers versatility in accepting images, audio, point cloud, and general signals, our tool stands out with its specialization in integrating game event data and providing an automatic feature report specifically designed for GBA scenarios. In addition, our tool allows to directly incorporate Unity projects into the environment, a capability challenging to achieve with generic annotation tools. Moreover, the *Videojot* tool [12] focuses on the annotation of video streams by combining zoom, drawing, and temporal social bookmarking, meanwhile Palotai's tool [13] emphasizes ML-based event recognition in video data to enable automated annotation. In contrast, our tool emphasizes manual annotation with customizable labels tailored for GBA scenarios, while also accepting multimedia data. Furthermore, Philbrick's *RIL-Contour* [14] is specifically designed for medical imaging datasets and promotes collaborative annotation by supporting concurrent multiuser workflows. While existing tools offer versatility, they lack specialized features required for effective labeling in GBA scenarios. Our tool addresses the specific needs of GBA research, enhancing the efficiency and accuracy of the data labeling process, and setting it apart from other generic or context-specific tools.

2. Software description

2.1. Software architecture

We have built the "GBA Labeling Tool" as a web application built using the Django framework. Our tool is developed as a server-side web application, providing robust and scalable capabilities. Users can access our tool using any web browser, making it easily accessible and compatible with any platform. In Fig. 1 we can see the complete platform's architecture.

Django is a web framework that uses Python for building very fast dynamic websites. It implements advanced security measures, including SQL injection prevention and data validation [15]. Django's

architecture is inherently scalable, enabling web applications to expand without causing any disruptions to their functioning and the flow of traffic [15]. Django offers scalability and speed optimization capabilities such as caching, load balancing, or horizontal scaling, making it an ideal choice for addressing these difficulties. Considering that the GBA Labeling Tool is expected to be used by a relatively small numbers of users concurrently, Django's scalability features align perfectly with our needs.

We employ the SQLite embedded database, ideal for most low to medium traffic applications, supporting tens of thousands of transactions per second and offering faster blob data processing compared to file systems [16]. The combination of SQLite's efficiency and Django's capabilities for improving performance of database queries ensures our tool can efficiently handle large datasets while maintaining responsiveness. Next, we present the different modules that form the architecture.

2.2. Data input

Users can upload *audio*, *video*, and *game event data*. Each individual replay must be linked to a game, user group (e.g., a classroom), and a user. If these details are not included along with the replays, the platform will use the default game, group, and/or user.

Multimedia data (audio and video) are directly stored as replays in the database. Regarding gameplay data, the tool incorporates a *custom parser* to transform the raw data, which can be received in multiple formats such as CSV (Comma-Separated Values) or JSON (JavaScript Object Notation), into a different format for further processing and visualization. Our parser maps every individual event in the log file into an *event* instance in the database. This event model includes essential information such as the timestamp, event type, and user-related details. In addition, our backend includes a *feature computing* module that analyzes the data by iterating through each individual event and calculates a set of useful features to add context to the labeled replay. There are two types of features: *context features*, related to the user up to that point in the gameplay, and *attempt features*, which are specific to that particular attempt. These features include information like the active and inactive time (in seconds), the number of completed levels, or a specific count of events.

2.3. Customization

Gameplay data can be fully customized by the user. The first customization option is event naming. Typically, event names defined by the original event model are lengthy or hard to read. By customizing each event's name, annotators can create more readable replays and make the annotation process more efficient.

Annotators can also define new types of events by combining existing ones along with a set of regular expression operators. Fig. 2 shows the "Custom events definition" page. On the left column of the page we can find the "items list", which includes the original events from a specific game, as well as a wildcard event (*Any*). On the right column, three operators are available: *?* matches the preceding event zero or one time; *** matches anything in place of the ***; and *+* matches the preceding event one or more times. To use these operators effectively and avoid errors or potential infinite loops, users should have some knowledge of regular expressions. The user can drag any of the items on the columns into the center box and combine them to create new events to be shown in textual replays. For instance, Fig. 2 shows the custom events definition interface, where a user is defining an event called *NFLAPS* as "(FLAP)+". This means that any sequence of one or more *FLAP* events will be replaced by the new custom event. By using custom events, annotators can adapt the replays to their specific needs, ensuring they are more representative of the gameplay.

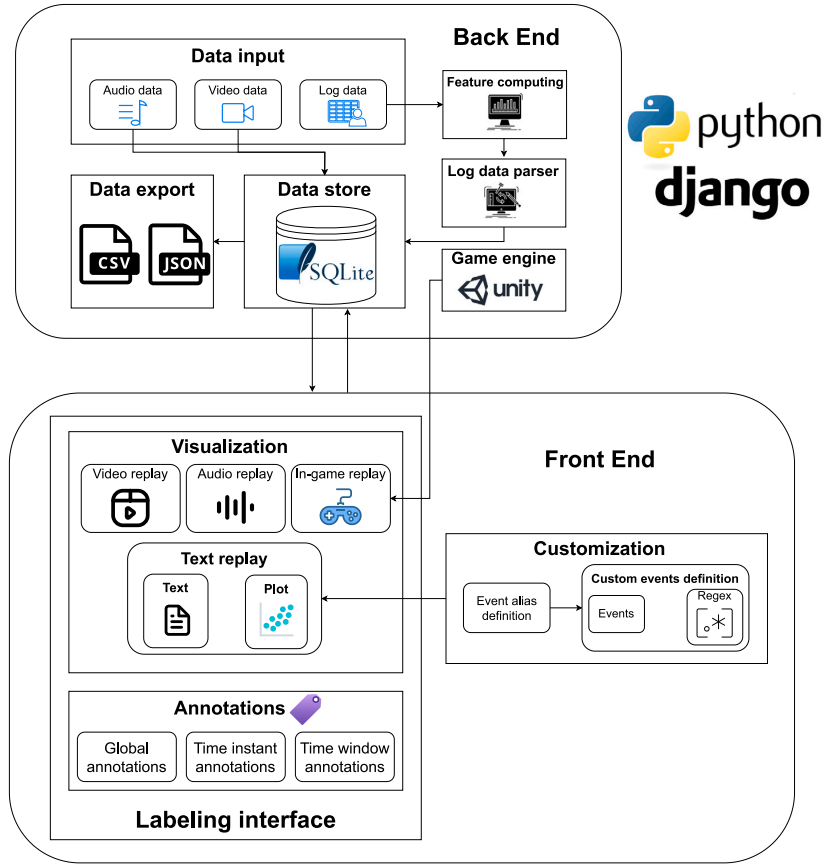


Fig. 1. Platform's architecture.

2.4. Visualization

The three different data formats result in four types of replays: *audio*, *video*, *in-game*, and *text* replays. When the user selects a specific replay, the tool loads the replay file to visualize it. For instance, in Fig. 3 we can see an *audio* replay, which is represented as waveforms.

We can visualize *game event data* in two different ways. Since log data is parsed and stored in the database, we can use the *game engine* itself to visualize the replay if the game allows it. In the example shown in 4(a), we use *Shadowspect*, a 3D geometry game designed as a formative assessment tool to measure math core standards [17]. The tool includes a set of templates that offer integration with *Unity WebGL* applications, simplifying the process for users to incorporate their projects into the tool environment.

Secondly, we can generate a textual (“pretty-printed”) representation using the original events generated by the users’ interaction with the games. In this example shown in Fig. 4(b), we can observe the game’s start, the number of level attempts, each action’s timestamp relative to the previous action, and the final outcome (completed or not). In both log data visualizations, features previously computed are incorporated in the right column. This data enrichment helps to provide a more informative and detailed understanding of replays.

In the text replay visualization interface, three different buttons located in the lower right corner offer annotators even more options to customize their data. The *Collapse/Uncollapse events* button allows merging consecutive identical actions, showing the number of times that the action has been performed consecutively. The *Replace custom events* button allows to replace the original events with their previously

defined custom events. Finally, the *Text/Visual mode* button allows users to create a plot using the log data to visually represent the events. For each replay, two plots are available: one generated with the original events and another generated using custom events. In Fig. 5 we can see an example of a plot generated using custom events and the same replay as in Fig. 4(b).

2.5. Annotations

In our tool, an annotation (or tag) is defined as the relationship between a user (annotator), a replay, a label, a value, a time interval, and a type of annotation. First, **global annotations** refer to the entire duration of the replay, indicating that a specific label value has been detected throughout the entire replay. Second, **time instant annotations** refer to a single point in time, indicating that a specific label value has been detected at a particular moment during the replay. Finally, **time window annotations** refer to a time interval (start and end) between the replay’s beginning and end.

When adding a time instant annotation to an audio or video replay, the tool will automatically assign the time instant to the moment displayed in the media player, as shown in Fig. 6(a). Moreover, when the user wants to add a time window annotation in an audio or video replay, they need to set the start and end of the time window using the respective buttons. For time instant annotations in a textual replay, a dropdown menu is available to select the event the user wants to annotate. Once selected, the tool assigns the global timestamp associated with that event to the annotation. Fig. 6(b) shows the annotation interface in a text replay when adding a time window annotation. We

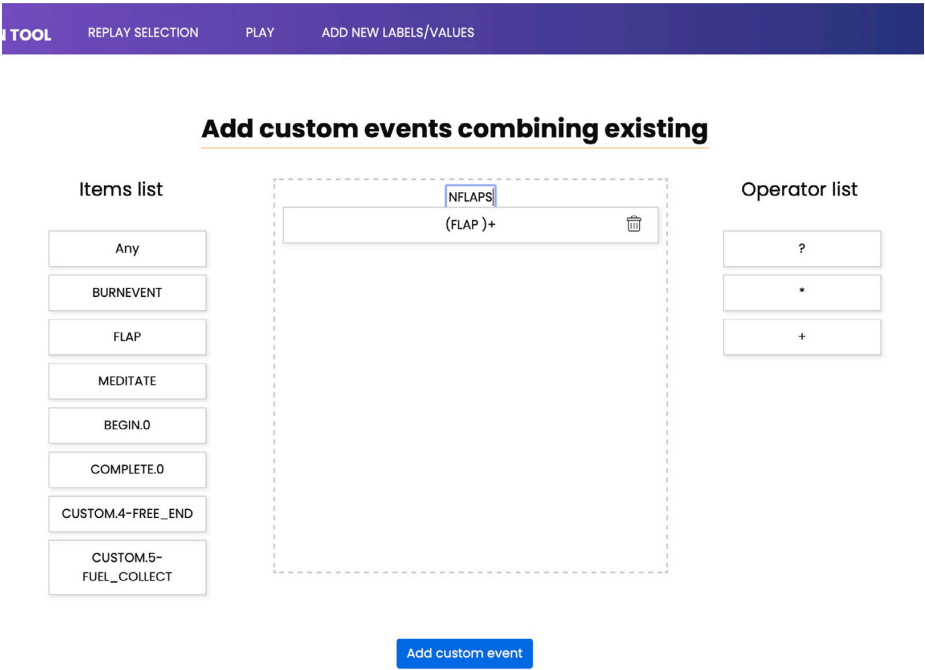


Fig. 2. Custom events definition page.



Fig. 3. Audio replay visualization.

can see a dropdown menu to select the type of annotation, event, label, and value, and two additional buttons to set the time window beginning and end.

2.6. Data export

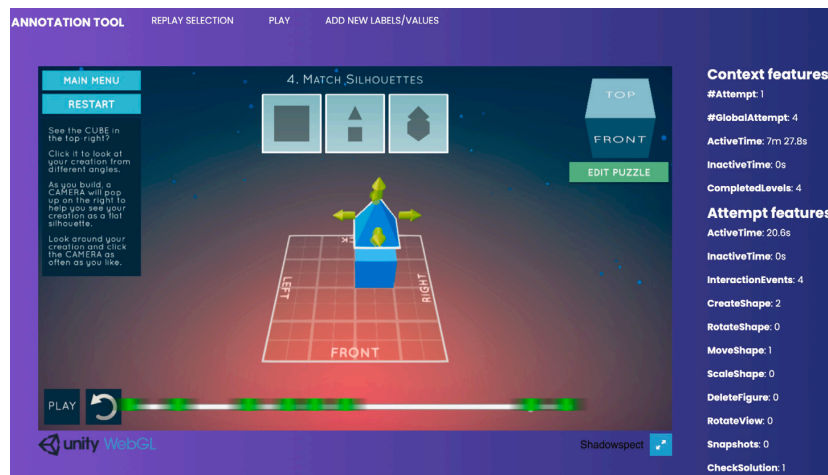
Users can export their annotated data at any time using the *Export* button available in the game, group, and user selection menus. The tool allows exporting data from all games or filtering data for specific games or groups. Two export formats are available: *JSON* and *CSV*, providing flexibility for users to choose the standard format that best suits their needs.

```
1 {
2   "label": "Creativity",
3   "typeTag": "TimeWindow",
```

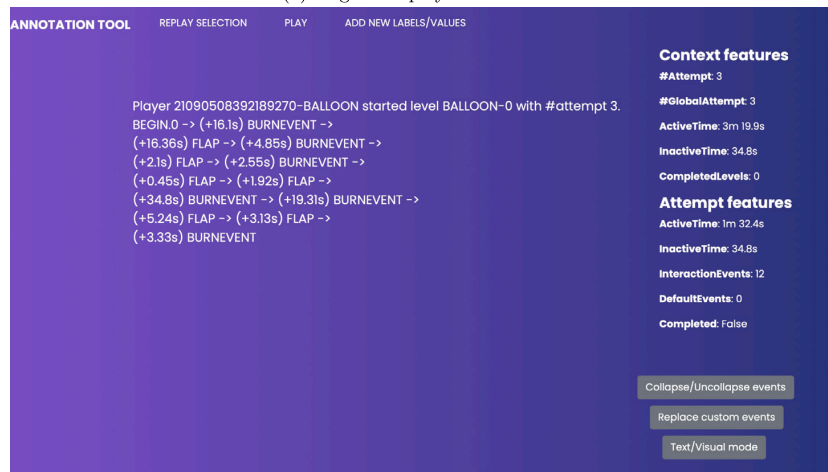
```
4   "value": "Creative",
5   "annotationTime": 7.4,
6   "finalAnnotationTime": 131.57,
7   "game": "BALLOON",
8   "user": 21090508420934984,
9   "level": "BALLOON-0",
10  "group": "MainGroup"
11 },...
```

Listing 1: Example of one of the JSON files generated by the export option.

Listing 1 shows a fragment of a JSON file generated by the export utility. As we can see, the JSON file contains all the information for each annotation added, including the type of tag, value, game, or level.



(a) In-game replay visualization.



(b) Text replay visualization.

Fig. 4. Log data visualizations.

3. Illustrative examples

To facilitate the reader's understanding and visualization, we have created a complementary video (<https://youtu.be/szOU9HL1QB0>) that demonstrates how a user would label game event data. Specifically, the user in this example aims to label the *persistence* competence by visualizing game replays from a game called *Crystal*.

First, the user begins by creating an alias for different game events and defining a custom event named "NRELEASES", defined as *one or more MOLECULERELEASE* events. Then, the user proceeds to label different replays by visualizing the game event data, collapsing events, replacing custom events, and plotting the data to obtain a better overview of each replay. Once the labeling process is complete, the video shows how the user exports the data into CSV format for downloading.

In addition, we have created another complementary video (<https://youtu.be/XhGKdZfutOk>) that provides a complete overview of the tool, including the use of audio, video, and game event data, as well as different annotation types.

4. Impact

The expected impacts of our tool primarily include the rise in adoption of multimedia and gameplay data labeling methodologies within GBA research. This tool is the first designed to meet specific requirements for annotating datasets in GBA scenarios, offering novel capabilities not found in existing annotation tools. We believe it can be applied in diverse GBA environments due to its versatility and adaptability to different types of multimedia data. This open-source tool contributes to the standardization of methodologies in GBA research, facilitating collaboration between researchers and the validation and comparison of results across studies.

One potential advantage is the reduction of annotation time, as the tool provides a user-friendly interface with customizable labeling options, helping practitioners to label data more efficiently. Moreover, the tool can contribute to higher consistency between annotations through predefined annotation types, labels, and values. Although using this labeling tool helps to optimize the annotation workflow, limitations still exist, as relying on manual annotations often results in noise and inconsistencies due to differences in annotators' interpretations of data and labeling criteria. However, by combining the strengths of our tool with training and consistency measures, researchers can mitigate

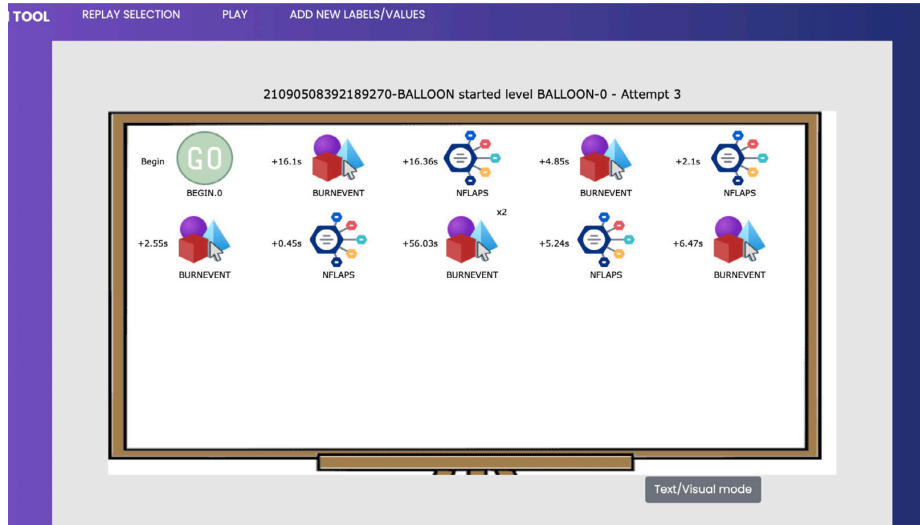
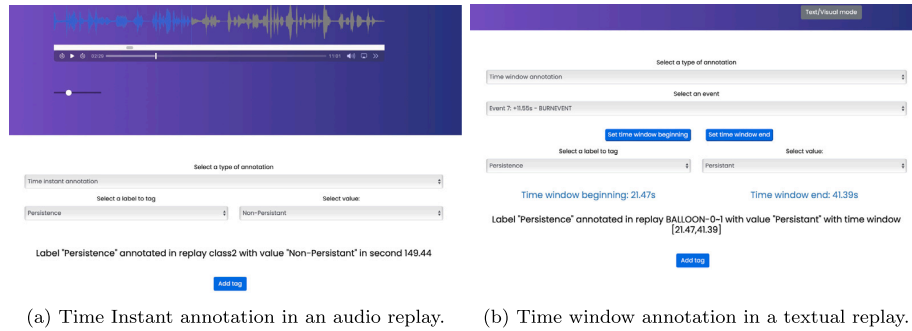


Fig. 5. Plot visualization using log data.



(a) Time Instant annotation in an audio replay.

(b) Time window annotation in a textual replay.

Fig. 6. Screenshots of the annotation interface in the tool.

potential issues and ensure the reliability of the annotated data. Moreover, exploring AI-assisted labeling techniques can help enhance the efficiency and accuracy of the annotation process, as well as enable the annotation of more complex and larger datasets.

Furthermore, the integration of annotated datasets with common analysis tools in the field allow researchers to conduct more complex analysis and gain deeper insights. There are several tools that are often used for statistical analysis and data visualization in educational research, such as *RapidMiner*, *Orange*, *SPSS*, and different packages in Python like *Scikit-learn* or *NumPy* [18,19]. Researchers can benefit from these tools and explore different correlations and patterns between game event data and learning outcomes, as well as apply sophisticated algorithms for predictive modeling or clustering, helping to advance the understanding of the impact of game-based learning on educational outcomes.

5. Conclusions

In this research, we introduce an open-source tool to address existing challenges in GBA data labeling. Our tool supports labeling of audio, video, and game event data, with a custom parser that integrates game event data to facilitate the analysis of gameplay performance and patterns. This includes a summary of game event data, as well as the possibility to define custom events by combining existing ones along with a set of regular expression operators. Moreover, users can employ

three annotation types and customize labels and values to meet the unique requirements of GBA scenarios. Additionally, the integration with Unity WebGL applications simplifies the process for users to incorporate their Unity projects into the labeling tool environment. Finally, users can export their labeled data in both CSV and JSON formats, facilitating data sharing and analysis.

The illustrative example demonstrates the practical potential of the proposed tool, and the open-source repository includes dataset samples that enable straightforward use of the tool. While our tool can help to optimize and standardize the annotation process, ensuring consistency among annotators may still require training and control. We plan to incorporate collaborative annotation features, such as calculating agreement between annotators, to reduce noisy data and improve annotations by identifying discrepancies. Furthermore, we plan to explore the integration of AI-assisted labeling techniques to enhance the efficiency and accuracy of the annotation process. This would help not only to reduce dependency on manual annotation but also to reduce noise between different annotators by providing automated assistance in the annotation process. Finally, to validate and enhance the tool's usability and accessibility, we plan to explore user feedback and conduct empirical usability studies, ensuring its accessibility to a wider audience and further validating its usability.

CRedit authorship contribution statement

Manuel J. Gomez: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **José A. Ruipérez-Valiente:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Félix J. García Clemente:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Manuel J. Gomez reports financial support was provided by Fundación Séneca. Manuel J. Gomez reports financial support was provided by Cybersecurity National Institute. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Sample data has been provided in the repository at <https://github.com/CyberDataLab/gba-labeling-tool/blob/main/sampleData>.

Acknowledgments

This work was partially supported by a) grant 21795/FPI/22 - Séneca Foundation. Cofinanced by Innovatiio Global Educación. Region of Murcia (Spain), b) REASSESS project (grant 21948/JLI/22) and SEMANTIC proof of concept (grant 22238/PDC/23), funded by the Seneca Foundation, Science and Technology Agency of the Region of Murcia, and c) the strategic project CDL-TALENTUM from the Spanish National Institute of Cybersecurity (INCIBE) and by the Recovery, Transformation and Resilience Plan, Next Generation EU.

References

- [1] Ullah M, Amin SU, Munsif M, Safaev U, Khan H, Khan S, et al. Serious games in science education. A systematic literature review. *Virtual Real Intell Hardw* 2022;4(3):189–209.
- [2] Becker K. What's the difference between gamification, serious games, educational games, and game-based learning. *Acad Lett* 2021;209:1–4.
- [3] Kato PM, de Klerk S. Serious games for assessment: Welcome to the jungle. *J Appl Test Technol* 2017;18(S1):1–6.
- [4] Smith SP, Blackmore K, Nesbitt K. A meta-analysis of data collection in serious games research. In: *Serious games analytics: Methodologies for performance measurement, assessment, and improvement*. Springer; 2015, p. 31–55.
- [5] Serrano Á, Marchiori EJ, del Blanco Á, Torrente J, Fernández-Manjón B. A framework to improve evaluation in educational games. In: *Proceedings of the 2012 IEEE global engineering education conference. EDUCON, IEEE*; 2012, p. 1–8.
- [6] Gao M, Zhang Z, Yu G, Arık SÖ, Davis LS, Pfister T. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In: *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part x 16*. Springer; 2020, p. 510–26.
- [7] Da Silva JL, Tabata AN, Broto LC, Cocron MP, Zimmer A, Brandmeier T. Open source multipurpose multimedia annotation tool. In: *Image analysis and recognition: 17th international conference, ICIAR 2020, póVoa de varzim, Portugal, June 24–26, 2020, proceedings, Part I 17*. Springer; 2020, p. 356–67.
- [8] Wang W, Xu G, Ding W, Huang GY, Li G, Tang J, et al. Representation learning from limited educational data with crowdsourced labels. *IEEE Trans Knowl Data Eng* 2020;34(6):2886–98.
- [9] Auer EM, Mersy G, Marin S, Blaik J, Landers RN. Using machine learning to model trace behavioral data from a game-based assessment. *Int J Sel Assess* 2022;30(1):82–102.
- [10] Chen F, Cui Y, Chu M-W. Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *Int J Artif Intell Educ* 2020;30:481–503.
- [11] Zhang Y, Wang Y, Zhang H, Zhu B, Chen S, Zhang D. Onelabeler: A flexible system for building data labeling tools. In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022, p. 1–22.
- [12] Riegler M, Lux M, Charvillat V, Carlier A, Vliegendhart R, Larson M. Videojot: A multifunctional video annotation tool. In: *Proceedings of international conference on multimedia retrieval*. 2014, p. 534–7.
- [13] Palotai Z, Láng M, Sárkány A, Tóser Z, Sonntag D, Toyama T, et al. Label-Movie: Semi-supervised machine annotation tool with quality assurance and crowd-sourcing options for videos. In: *2014 12th international workshop on content-based multimedia indexing. CBMI, IEEE*; 2014, p. 1–4.
- [14] Philbrick KA, Weston AD, Akkus Z, Kline TL, Korfiatis P, Sakinis T, et al. RIL-contour: A medical imaging dataset annotation tool for and with deep learning. *J Digit Imaging* 2019;32:571–81.
- [15] Chen S, Ahmmed S, Lal K, Deming C. Django web development framework: Powering the modern web. *Am J Trade Policy* 2020;7(3):99–106.
- [16] Gaffney KP, Prammer M, Brasfield L, Hipp DR, Kennedy D, Patel JM. Sqlite: past, present, and future. *Proc VLDB Endow* 2022;15(12).
- [17] Arcade ME. *ShadowSpect*. 2020, <https://education.mit.edu/project/shadowspect/>. [Online; Last Accessed 15 September 2023].
- [18] Purwanto A. Education research quantitative analysis for little respondents: comparing of lisrel, tetrad, GSCA, amos, SmartPLS, WarpPLS, and SPSS. *J Studi Guru Dan Pembelajaran* 2021;4(2).
- [19] Slater S, Joksimović S, Kovanovic V, Baker RS, Gasevic D. Tools for educational data mining: A review. *J Educ Behav Stat* 2017;42(1):85–106.

Conclusions and future directions

This section presents several key conclusions and highlights future directions emerging from the completion of this Ph.D. thesis and its associated research findings.

C1. The need for more standard assessment frameworks

This Ph.D. thesis demonstrates that GBA is a powerful tool for extracting valuable knowledge from user data. Such data can be collected from various sources, including audio and video recordings, log-event data generated from learners' interactions, and multimodal inputs such as gestures, facial expressions, or biometric signals. However, although large data repositories are increasingly being created and made available for analysis, result *R1* evidenced the lack of standardized frameworks for both designing and implementing assessments that effectively leverage these rich datasets. This thesis introduces a modular and scalable solution for handling GBA data using an ontology-based architecture, as demonstrated in results *R2* and *R3*. However, there remains a need for researchers to adopt standard data formats instead of relying on specific assessment machinery. Standardization facilitates the open sharing of data for other research purposes and reduces the effort required to replicate results using similar techniques.

Moreover, as the number and size of these datasets continue to grow, there is an increasing need for efficient processing systems capable of processing large-scale data, such as the one presented in result *R3*. Particularly in the era of Big Data, developing a scalable and efficient architecture for GBAs is essential.

C2. The increasing value of GBA for education and training

GBAs have the potential to be applied in a wide range of contexts and situations. Result *R1* revealed that these assessments are mainly used in educational settings, but also in medical environments for purposes such as rehabilitation, and in professional environments. In the latter context, companies use games not only for staff recruitment but also to evaluate employee performance and provide additional feedback. Regarding the use of GBAs in education, they are most commonly applied in middle and high school, as children and adolescents are ideal target users due to their familiarity with gaming environments and mechanics. A particularly promising aspect of these assessments in education is their potential to support

the development and sustainability of 21-st century skills, such as collaboration, communication, and persistence. These skills are traditionally difficult to measure using conventional assessment methods, and benefit from being applied in context for more accurate measurements. However, existing literature still reports a lack of research focused on developing and measuring 21-st century skills [35].

For games effectively assess complex skills and behaviors, they must be grounded in strong game design principles, adopting design-based research methods. Currently, further research is needed to systematically develop and enhance the current design of games for assessment purposes, as many existing examples still use simple quizzes, either as their primary assessment mechanism or as a significant component of the gameplay.

C3. Complexity in skills demands complexity in assessment

One of the key potentials of GBA is its ability to measure complex cognitive skills and behaviors, as it allows for the recreation of more authentic and realistic scenarios required to assess the application of these skills in context. The systematic review in result *R1* evidenced the need for more sophisticated assessment methods, since researchers often rely on simple metrics and indicators that fail to capture the complexity and context-dependence nature of cognitive skills. There is a variety of methods that can help us with these challenges, including ML, DL, knowledge inference, and data science techniques such as sequence and pattern mining. These approaches focus on discovering hidden patterns and behavioral sequences that are not immediately evident, providing valuable information into learning progress and skill development. Moreover, the integration of multimodal data, such as physiological signals, audio, and video recordings, can broaden the scope of assessments by providing a more holistic understanding of user interactions and states.

However, employing AI models introduces an additional layer of difficulty. These models often require large datasets with consistent labeling to perform well, and such data can be difficult and costly to obtain. In this context, result *R5* presents a practical tool that offers a pathway towards more accurate and efficient data labeling in the GBA context, supporting the annotation of both multimedia and log-event data within a single platform. Of course, the design of meaningful assessment strategies is crucial to ensure that this labeling process captures relevant and valuable information.

C4. Interpretability as a key enabler for real-world application

As new AI models and methods emerge, there is an increasing demand from researchers and stakeholders for more understandable and transparent outcomes. The performance improvement of these methods is usually achieved through increased model complexity, which turn the developed system into a “black box.” This need is particularly important given the perception among non-technical users, who often see AI as producing outcomes that are inexplicable or difficult to interpret. As a result, research interest in the field of XAI, which focuses on developing methods

to explain and interpret AI models, has experienced a significant growth in recent years.

In the context of GBA, key stakeholders such as teachers, corporate trainers, or medical instructors, are often non-technical users. In practice, it is frequently observed that final users find difficult to interpret even basic metrics and indicators. This challenge is often addressed through the use of dashboards and visualization techniques, which graphically represent the data in a more accessible and intuitive manner. However, emerging techniques require the application of XAI to enable appropriate interpretation of complex model outputs. In this thesis, result *R4* addresses the interpretability of performance prediction models by identifying and explaining the features that contributed to the model's prediction of whether a learner would complete a task. This empowers teachers with a clear understanding of the model's reasoning and provides the opportunity to analyze learning progress and intervene if necessary.

From this Ph.D. thesis and its results, four clear avenues are opened to explore as a continuation of this GBA research line.

F1. Frameworks for scalable design and integration of GBAs

The results of this thesis reinforce the critical role of well-grounded designs in providing valid and meaningful assessments. A key area of future work involves the development of frameworks that systematically guide the design of GBAs. This may include domain-specific games adapted to emerging educational trends, such as fostering AI literacy or combating disinformation. These frameworks should support explicit mappings between competencies to be assessed, methodologies, game mechanics, and the data to be collected, therefore enhancing both validity and scalability.

As part of these frameworks, it is crucial to align game concepts with existing curricula to facilitate their adoptions by educators, since teachers are still unsure about how to integrate game activities with the regular curriculum. This underscores the importance of developing clear implementation strategies that guide how GBAs can be used effectively in real-world settings. Bridging educational goals with GBA mechanics through these frameworks would contribute to its broader adoption and impact across diverse educational contexts.

F2. Deployment of GBAs solutions in real environments

One of the limitations of the interoperable assessment framework developed in this work is the lack of validation in real environments. In fact, the validation challenge is commonly reported in the literature, with many studies focusing on the technical implementation without evaluating the validity or alignment of their metrics with learning outcomes. Future research could address this gap by deploying GBAs solutions in real-world environments, enabling the analysis of their validity through external measures, as well as their usability and accessibility.

This future research direction directly affects the work conducted in this Ph.D. thesis: although the metrics and indicators selected for implementation in the interoperable architecture were extracted from previous research, the study did not conduct validity or reliability tests to ensure that the reported assessments were trustworthy. Addressing this limitation would strengthen the credibility and applicability of the proposed framework in real-world settings. Validation at multiple levels, ranging from learning outcome alignment to practical considerations such as user experience, accessibility, and contextual relevance, would enhance the framework's robustness and impact.

F3. Human-in-the-Loop (HITL) GBA approaches

The practical labeling tool developed in this dissertation paves the way for HITL methodologies in the context of GBA, as this tool enables human experts to interact with data, provide annotations, and iteratively refine assessment models. First, future work could leverage the tool for exploring methodologies that capture complex patterns from labeled data, including the use of more traditional approaches such as ML, as well as emerging techniques like sequence mining, temporal pattern recognition, or semi-supervised learning. These methods could support the development of sophisticated assessment models capable of inferring knowledge acquisition, cognitive strategies, or skill development. For example, a case study could cover an assessment of domain-specific knowledge like mathematics or history, while another could explore the evaluation of skills like collaboration or creativity.

In parallel, an important aspect to consider is the usability of the tool itself. Future work could evaluate the usability and efficiency of the labeling tool from the perspective of annotators and researchers. Comparative studies could also be conducted to assess the effectiveness of the different visualization methods provided in the tool, examining which approaches are more efficient in terms of labeling time and which are better to detect finer details or nuanced patterns in the labeled data.

F4. Multimodal GBAs

The majority of the data used in this research corresponds to log data generated from player interactions with games. However, GBAs can benefit from incorporating many different data sources, such as physiological signals (e.g., heart rate, skin conductance), eye-tracking data, or audio and video recordings. For instance, the incorporation of physiological signals is already being explored in different rehabilitation contexts like upper-limb movement recovery, or to assess cognitive functions such as attention, memory, and executive control. In addition, the combination of log data with multimodal data provides access to implicit user states like attention or stress, which could allow for fine-grained models by offering converging evidence from multiple sources of data.

Therefore, future work should explore multimodal GBA approaches, considering not only technical aspects to process heterogeneous data types, but also methodological frameworks to model the relationship between observable behaviors and internal cognitive or emotional states.

Bibliography

- [1] N. Selwyn, *Education and Technology: Key Issues and Debates*, English, 2nd. United Kingdom: Bloomsbury Academic, 2017, ISBN: 9781474235914.
- [2] R. E. Clark, “Learning from serious games? arguments, evidence, and research suggestions”, *Educational Technology*, vol. 47, no. 3, pp. 56–59, 2007.
- [3] F. Laamarti, M. Eid, and A. El Saddik, “An overview of serious games”, *International Journal of Computer Games Technology*, vol. 2014, no. 1, p. 358 152, 2014.
- [4] Entertainment Software Association, *2024 essential facts about the u.s. video game industry*, Accessed: 2025-04-02, 2024. [Online]. Available: <https://www.theesa.com/resources/essential-facts-about-the-us-video-game-industry/2024-data/>.
- [5] Video Games Europe, *2023 video games - european key facts*, Accessed: 2025-04-02, 2023. [Online]. Available: <https://www.videogameseurope.eu/publication/2023-video-games-european-key-facts/>.
- [6] R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer, *Serious games*. Springer, 2016.
- [7] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth, “Digital games, design, and learning: A systematic review and meta-analysis”, *Review of educational research*, vol. 86, no. 1, pp. 79–122, 2016.
- [8] S. al-Qallawi and M. Raghavan, “A review of online reactions to game-based assessment mobile applications”, *International Journal of Selection and Assessment*, vol. 30, no. 1, pp. 14–26, 2022.
- [9] J. Wiemeyer and A. Kliem, “Serious games in prevention and rehabilitation—a new panacea for elderly people?”, *European Review of Aging and Physical Activity*, vol. 9, pp. 41–50, 2012.
- [10] P. M. Kato and S. de Klerk, “Serious games for assessment: Welcome to the jungle”, *Journal of Applied Testing Technology*, pp. 1–6, 2017.
- [11] A. N. Rafferty, M. M. LaMar, and T. L. Griffiths, “Inferring learners’ knowledge from their actions”, *Cognitive Science*, vol. 39, no. 3, pp. 584–618, 2015.
- [12] V. J. Shute and S. Rahimi, “Stealth assessment of creativity in a physics video game”, *Computers in Human Behavior*, vol. 116, p. 106 647, 2021.

- [13] V. Shute and M. Ventura, “Stealth assessment”, *The SAGE encyclopedia of educational technology*, pp. 675–676, 2015.
- [14] K. E. DiCerbo, “Game-based assessment of persistence”, *Journal of Educational Technology & Society*, vol. 17, no. 1, pp. 17–28, 2014.
- [15] M. Qian and K. R. Clark, “Game-based learning and 21st century skills: A review of recent research”, *Computers in human behavior*, vol. 63, pp. 50–58, 2016.
- [16] M. Freire, Á. Serrano-Laguna, B. Manero Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, “Game learning analytics: Learning analytics for serious games”, in *Learning, design, and technology: An international compendium of theory, research, practice, and policy*, Springer, 2023, pp. 3475–3502.
- [17] X. Ge and D. Ifenthaler, “Designing engaging educational games and assessing engagement in game-based learning”, in *Gamification in education: Breakthroughs in research and practice*, IGI global, 2018, pp. 1–19.
- [18] Y. J. Kim and D. Ifenthaler, “Game-based assessment: The past ten years and moving forward”, *Game-based assessment revisited*, pp. 3–11, 2019.
- [19] Á. Serrano-Laguna, I. Martínez-Ortiz, J. Haag, D. Regan, A. Johnson, and B. Fernández-Manjón, “Applying standards to systematize learning analytics in serious games”, *Computer Standards & Interfaces*, vol. 50, pp. 116–123, 2017.
- [20] C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, “Applications of data science to game learning analytics data: A systematic literature review”, *Computers & Education*, vol. 141, p. 103612, 2019.
- [21] M. Frutos-Pascual and B. G. Zapirain, “Review of the use of ai techniques in serious games: Decision making and machine learning”, *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 9, no. 2, pp. 133–152, 2015.
- [22] J. A. Caballero Hernández, M. Palomo Duarte, J. M. Dodero Beardo, D. Gašević, *et al.*, “Supporting skill assessment in learning experiences based on serious games through process mining techniques”, *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 6, pp. 146–159, 2024.
- [23] M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “A systematic literature review of game-based assessment studies: Trends and challenges”, *IEEE Transactions on Learning Technologies*, vol. 16, no. 4, pp. 500–515, 2022.
- [24] M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “Developing and validating interoperable ontology-driven game-based assessments”, *Expert Systems with Applications*, vol. 248, p. 123370, 2024.
- [25] M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. García Clemente, “A framework to support interoperable game-based assessments as a service (gbaaas): Design, development, and use cases”, *Software: Practice and Experience*, vol. 53, no. 11, pp. 2222–2240, 2023.

-
- [26] M. J. Gomez, Á. Armada Sánchez, M. Albaladejo-González, F. J. García Clemente, and J. A. Ruipérez-Valiente, “Utilising explainable ai to enhance real-time student performance prediction in educational serious games”, *Expert Systems*, vol. 42, no. 3, e70008, 2025.
- [27] M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, “Optimizing multimedia and gameplay data labeling: A web-based tool for game-based assessment”, *SoftwareX*, vol. 27, p. 101763, 2024.
- [28] F. Bellotti, B. Kapralos, K. Lee, P. Moreno-Ger, and R. Berta, “Assessment in and of serious games: An overview”, *Advances in Human-Computer Interaction*, vol. 2013, no. 1, p. 136864, 2013.
- [29] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai”, *Information fusion*, vol. 58, pp. 82–115, 2020.
- [30] M. Gao, Z. Zhang, G. Yu, S. Ö. Arik, L. S. Davis, and T. Pfister, “Consistency-based semi-supervised active learning: Towards minimizing labeling cost”, in *European Conference on Computer Vision*, Springer, 2020, pp. 510–526.
- [31] M. J. Page, D. Moher, P. M. Bossuyt, *et al.*, “Prisma 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews”, *bmj*, vol. 372, 2021.
- [32] M. Fernández López, A. Gómez-Pérez, and N. Juristo Juzgado, “Methontology: From ontological art towards ontological engineering”, in *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series*, Ontology Engineering Group - OEG, American Association for Artificial Intelligence, 1997.
- [33] J. Lehmann, G. Sejdiu, L. Bühmann, *et al.*, “Distributed semantic analytics using the sansa stack”, in *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*, Springer, 2017, pp. 147–155.
- [34] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions”, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777.
- [35] M. H. Hussein, S. H. Ow, M. M. Elaish, and E. O. Jensen, “Digital game-based learning in k-12 mathematics education: A systematic literature review”, *Education and Information Technologies*, vol. 27, no. 2, pp. 2859–2891, 2022.