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

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TITULO 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 (GBAaaS). 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, GBAaaS). 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

Chapter 1

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]

Chapter 2

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.

Chapter 3

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.

Chapter 4

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 SANSA-Stack 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.

Chapter 5

Publications

1 GBA: Current Trends and Challenges

Title

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

Authors

Manuel J. Gomez[1], José A. Ruipérez-Valiente[1], Félix J. García Clemente[1]

[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.

2 Towards Semantic Interoperability

Title

Developing and Validating Interoperable Ontology-driven Game-Based Assessments

Authors

Manuel J. Gomez[1], José A. Ruipérez-Valiente[1], Félix J. García Clemente[1]

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

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

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).

3 A Framework for Interoperable GBAaaS

Title

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

Authors

Manuel J. Gomez[1], José A. Ruipérez-Valiente[1], Félix J. García Clemente[1]

[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.

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. Gomez[1],Álvaro Armada Sánchez[1],
Mariano Albaladejo-González[1], Félix J. García Clemente[1],
José A. Ruipérez-Valiente[1]

[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.

5 Optimizing Manual Labeling in GBA

Title

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

Authors

Manuel J. Gomez[1], José A. Ruipérez-Valiente[1], Félix J. García Clemente[1]

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

Publication details

Journal	SoftwareX	**Publisher**	Elsevier
Volume	27	**Number**	–
Pages	101763	**Year**	2024
JIF	2.4	**Rank**	Q2
Status	Published	**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.

Chapter 6

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 a 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 on 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. 103 612, 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. Gaševic, *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. 123 370, 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. 101 763, 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. 136 864, 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. Arık, 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 Asociation 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.