

Argumento de validez para la evaluación del aprendizaje en entornos inmersivos en la docencia universitaria

A Validity Argument for the Assessment of Learning in Immersive Environments in University Teaching

Un argomento di validità per la valutazione dell'apprendimento negli ambienti immersivi nella docenza universitaria

Elisa Flor Guamán Troya¹

Email: fguamant@unemi.edu.ec

ORCID: <https://orcid.org/0009-0005-3898-5161>

Michelle Giovanna Ortiz Ripalda²

Email: mortizr@unemi.edu.ec

ORCID: <https://orcid.org/0009-0003-0675-6001>

Andrea Lisette Vargas Arias³

Email: avargasa@unemi.edu.ec

ORCID: <https://orcid.org/0009-0009-3658-4486>

Correspondencia: Elisa Flor Guamán Troya¹ **Email:** fguamant@unemi.edu.ec

Artículo de revisión

* **Recibido:** 15-09-2026 * **Aceptado:** 30-09-2026 * **Publicado:** 23-10-2026

1. Facultad de Posgrado, Universidad Estatal de Milagro (UNEMI), Milagro, Ecuador.
2. Facultad de Posgrado, Universidad Estatal de Milagro (UNEMI), Milagro, Ecuador.
3. Facultad de Posgrado, Universidad Estatal de Milagro (UNEMI), Milagro, Ecuador.

Resumen

Las universidades incorporan de forma creciente la realidad virtual, la realidad aumentada, la realidad mixta y los mundos virtuales tridimensionales para que el estudiantado demuestre competencias en escenarios cercanos a la práctica profesional. Esta expansión ha sido más rápida que la reflexión sobre la validez de las calificaciones obtenidas en esos entornos: cuando el desempeño está mediado por una interfaz, un avatar y unas condiciones técnicas variables, cambian las fuentes de error y las inferencias que sostienen la decisión docente. El objetivo de este artículo fue proponer un argumento de validez para la evaluación del aprendizaje en entornos inmersivos en la docencia universitaria, fundamentado en el enfoque argumentativo de Kane y en la teoría unificada de Messick. Se realizó una revisión teórica crítica de 41 fuentes cuya existencia y metadatos se verificaron en la fuente editorial, organizadas en siete categorías: teoría de la validez y estándares, validación en simulación, diseño y modelos de medición, evaluación auténtica, aprendizaje inmersivo, factores del medio, y ética, equidad y normativa. El estudio no recogió datos de campo. El resultado es un argumento interpretativo y de uso con cuatro inferencias (puntuación, generalización, extrapolación y decisión), cada una con sus supuestos, las amenazas propias del medio inmersivo y las evidencias necesarias para respaldarla. Se identificaron como amenazas principales la varianza irrelevante al constructo derivada de la familiaridad tecnológica, del malestar por simulador y de la carga cognitiva extrínseca; la variabilidad entre escenarios y ocasiones; los sesgos del evaluador ante desempeños mediados por avatares; y las desigualdades de acceso. Se propone un plan de validación en cuatro fases aplicable por equipos docentes. Se concluye que la autenticidad del escenario no garantiza la validez de la calificación y que esta debe argumentarse y comprobarse antes de usar las puntuaciones para decisiones de alto impacto.

Palabras clave: validez, evaluación del aprendizaje, realidad virtual, entornos inmersivos, educación superior.

Abstract

Universities are increasingly incorporating virtual reality, augmented reality, mixed reality and three-dimensional virtual worlds so that students can demonstrate competences in scenarios close to professional practice. This expansion has outpaced reflection on the validity of the scores obtained in these environments: when performance is mediated by an interface, an avatar and variable technical conditions, the sources of error and the inferences supporting the teacher's decision change. This article aimed to propose a validity argument for the assessment of learning in immersive environments in university teaching, grounded in Kane's argument-based approach and Messick's unified theory. A critical theoretical review of 41 sources, whose existence and metadata were verified at the publisher, was carried out and organized into seven categories: validity theory and standards, validation in simulation, design and measurement models, authentic assessment, immersive learning, medium-related factors, and ethics, equity and

regulation. No field data were collected. The result is an interpretation and use argument with four inferences (scoring, generalization, extrapolation and decision), each with its assumptions, the threats specific to the immersive medium and the evidence needed to support it. The main threats identified were construct-irrelevant variance arising from technological familiarity, simulator sickness and extraneous cognitive load; variability across scenarios and occasions; rater bias towards avatar-mediated performances; and unequal access. A four-phase validation plan applicable by teaching teams is proposed. It is concluded that the authenticity of a scenario does not guarantee the validity of the score, which must be argued and tested before scores are used for high-stakes decisions.

Keywords: validity, assessment of learning, virtual reality, immersive environments, higher education.

Riassunto

Le università integrano sempre più spesso la realtà virtuale, la realtà aumentata, la realtà mista e i mondi virtuali tridimensionali affinché gli studenti dimostrino competenze in scenari vicini alla pratica professionale. Questa espansione è stata più rapida della riflessione sulla validità dei punteggi ottenuti in tali ambienti: quando la prestazione è mediata da un'interfaccia, da un avatar e da condizioni tecniche variabili, cambiano le fonti di errore e le inferenze che sostengono la decisione del docente. L'obiettivo dell'articolo è stato proporre un argomento di validità per la valutazione dell'apprendimento negli ambienti immersivi nella docenza universitaria, fondato sull'approccio argomentativo di Kane e sulla teoria unitaria di Messick. È stata condotta una revisione teorica critica di 41 fonti, di cui sono stati verificati esistenza e metadati presso l'editore, organizzate in sette categorie. Non sono stati raccolti dati sul campo. Il risultato è un argomento interpretativo e d'uso con quattro inferenze (attribuzione del punteggio, generalizzazione, estrapolazione e decisione), ciascuna con i propri presupposti, le minacce specifiche del mezzo immersivo e le evidenze necessarie. Le minacce principali sono la varianza estranea al costrutto dovuta alla familiarità tecnologica, al malessere da simulatore e al carico cognitivo estraneo; la variabilità tra scenari e occasioni; i pregiudizi del valutatore verso prestazioni mediate da avatar; e le disuguaglianze di accesso. Si propone un piano di validazione in quattro fasi. Si conclude che l'autenticità dello scenario non garantisce la validità del punteggio.

Parole chiave: validità, valutazione dell'apprendimento, realtà virtuale, ambienti immersivi, istruzione superiore.

Introducción

La docencia universitaria atraviesa una ampliación de los escenarios en los que el estudiantado aprende y demuestra lo aprendido. A las aulas presenciales, a los laboratorios y a las plataformas virtuales se suman hoy entornos inmersivos: simulaciones de realidad virtual que se recorren con un visor, aplicaciones de realidad aumentada que superponen información digital sobre objetos físicos, sistemas de realidad mixta y mundos virtuales tridimensionales en los que docentes y estudiantes interactúan mediante avatares. Milgram y Kishino (1994) describieron estas modalidades como puntos de un continuo que va del entorno real al completamente virtual, y en este trabajo se utiliza la expresión realidades híbridas para referirse a ese continuo y a sus combinaciones. Lo que tienen en común es la capacidad de producir presencia, es decir, la sensación de estar en un lugar y de que lo que ocurre es real, lo que lleva a las personas a responder de manera semejante a como lo harían en la situación auténtica (Slater y Sanchez-Vives, 2016).

Las síntesis disponibles muestran que estas tecnologías ya forman parte de la educación superior. Radianti et al. (2020) identificaron aplicaciones de realidad virtual inmersiva en ingeniería, ciencias de la salud, ciencias naturales, arquitectura y formación docente, y advirtieron que la mayoría de los estudios carecía de fundamentación teórica y de un diseño de evaluación riguroso. Los metaanálisis sobre realidad virtual (Merchant et al., 2014; Wu et al., 2020), realidad aumentada (Garzón y Acevedo, 2019) y aprendizaje basado en simulación (Chernikova et al., 2020) informan efectos positivos sobre el aprendizaje, aunque heterogéneos y dependientes del diseño instruccional. El interés por los mundos virtuales compartidos se ha renovado bajo la noción de metaverso (Mystakidis, 2022), especialmente en programas en línea y de posgrado, donde la distancia física limita las evaluaciones presenciales.

La investigación se ha centrado casi siempre en si estos entornos mejoran el aprendizaje, y mucho menos en si las calificaciones que se asignan dentro de ellos significan lo que el docente supone que significan. Hamilton et al. (2021) señalaron que buena parte de los estudios sobre realidad virtual mide los resultados con instrumentos externos a la simulación y con diseños experimentales débiles. Esta omisión tiene consecuencias prácticas: cuando un docente califica a un estudiante por lo que hizo en un laboratorio virtual o por su intervención en una sala del metaverso, toma una decisión sobre su competencia. Si el desempeño estuvo condicionado por la falta de práctica con el visor, por el mareo, por una conexión inestable o por la apariencia del avatar, la calificación no refleja solo el aprendizaje y la decisión puede ser injusta.

La teoría contemporánea de la validez ofrece herramientas para abordar este problema. Messick (1995) definió la validez como un juicio evaluativo integrado sobre el grado en que la evidencia empírica y los fundamentos teóricos respaldan la pertinencia de las interpretaciones y de los usos de las puntuaciones. Desde esta perspectiva, la validez no es una propiedad del instrumento, sino de las interpretaciones y usos que se hacen de sus resultados, y su análisis incluye las consecuencias. Kane (2006, 2013) tradujo esa concepción en un enfoque

argumentativo: primero se explicita un argumento interpretativo y de uso que encadena las inferencias necesarias para pasar del desempeño observado a la decisión, y después se reúnen evidencias para comprobar la plausibilidad de cada eslabón. Los estándares profesionales de medición educativa adoptan esta misma lógica (American Educational Research Association et al., 2014).

El enfoque argumentativo se ha aplicado con éxito a la evaluación mediante simulación en la educación médica (Cook et al., 2015; Cook y Hatala, 2016), pero no se ha desarrollado un argumento de validez específico para la evaluación del aprendizaje en entornos inmersivos en la docencia universitaria general, ni en el contexto latinoamericano. Este vacío es relevante en Ecuador, donde el Reglamento de Régimen Académico concede a las instituciones de educación superior autonomía para definir sus sistemas de evaluación (Consejo de Educación Superior, 2022) y donde la expansión de la modalidad en línea en el posgrado ha impulsado la búsqueda de evaluaciones más auténticas.

En este marco, el estudio se orientó por la pregunta: ¿qué argumento de validez debe sostener el uso de las calificaciones obtenidas en entornos inmersivos para tomar decisiones sobre el aprendizaje en la docencia universitaria? El objetivo general fue proponer un argumento interpretativo y de uso para la evaluación del aprendizaje en entornos inmersivos, fundamentado en la teoría de la validez. Los objetivos específicos fueron: (a) caracterizar los rasgos de la evaluación en entornos inmersivos que afectan la validez; (b) formular las inferencias del argumento y sus supuestos; (c) identificar las amenazas a la validez propias del medio; y (d) proponer un plan de evidencias aplicable por equipos docentes universitarios.

Metodología

Tipo de estudio

Se realizó una revisión teórica crítica con orientación propositiva. Este tipo de estudio no busca estimar efectos ni agregar resultados empíricos, como haría una revisión sistemática o un metaanálisis, sino integrar marcos teóricos y evidencias de distintas áreas para construir una propuesta conceptual fundamentada. El producto esperado fue un argumento de validez, entendido en el sentido de Kane (2013) como una cadena explícita de inferencias y supuestos que puede someterse a comprobación empírica. El trabajo no recogió datos de campo ni involucró a participantes humanos.

Selección y verificación de las fuentes

Las fuentes se seleccionaron de manera intencional según su pertinencia para cada componente del argumento. Se priorizaron, en primer lugar, los textos fundacionales de la teoría de la validez y de la medición del desempeño; en segundo lugar, las síntesis de investigación (revisiones sistemáticas y metaanálisis) sobre aprendizaje inmersivo y basado en simulación en la educación superior; y en tercer lugar, los estudios que describen factores del medio capaces de

alterar el desempeño, como la presencia, la carga cognitiva, el malestar por simulador y la representación mediante avatares. Se incluyeron además los documentos normativos ecuatorianos pertinentes. Como criterios de inclusión se exigió que la fuente fuera identificable y accesible, que estuviera publicada en una revista arbitrada, en una editorial académica o por un organismo oficial, y que su contenido aportara a alguno de los componentes del argumento. Se excluyeron los textos de divulgación, los materiales comerciales y las fuentes cuyos metadatos no pudieron verificarse. La existencia, la autoría, el año, la revista o editorial y, cuando correspondía, el DOI de cada fuente se contrastaron en el sitio de la editorial o en repositorios institucionales. El corpus final quedó formado por 41 fuentes, distribuidas en las categorías que se presentan en la Tabla 1.

Tabla 1. Composición del corpus de fuentes analizadas

Categoría	n	Fuentes
Teoría de la validez y estándares	9	American Educational Research Association et al. (2014); Downing (2003); Downing y Haladyna (2004); Kane (2006, 2013); Kane et al. (1999); Linn et al. (1991); Messick (1995); Moss (1994).
Validación de la evaluación en simulación	2	Cook et al. (2015); Cook y Hatala (2016).
Diseño de la evaluación y modelos de medición	4	Brennan (2001); Linacre (1989); Mislevy et al. (2003); Shute (2011).
Evaluación auténtica y juicio evaluativo	3	Gulikers et al. (2004); Sadler (1989); Villarroel et al. (2018).
Aprendizaje inmersivo y basado en simulación	12	Chernikova et al. (2020); Garzón y Acevedo (2019); Hamilton et al. (2021); Jensen y Konradsen (2018); Makransky et al. (2019); Makransky y Petersen (2021); Merchant et al. (2014); Milgram y Kishino (1994); Mystakidis (2022); Radianti et al. (2020); Slater y Sanchez-Vives (2016); Wu et al. (2020).
Factores del medio que afectan el desempeño	5	Bailenson (2021); Kennedy et al. (1993); Sweller (1988); Witmer y Singer (1998); Yee y Bailenson (2007).
Ética, equidad y normativa	6	Asamblea Nacional del Ecuador (2021); CAST (2018); Consejo de Educación Superior (2022); O'Brolcháin et al. (2016); Raes et al. (2020); Slater et al. (2020).
Total	41	

Nota. Las 41 fuentes coinciden con la lista de referencias. Se contabiliza por separado cada obra de un mismo autor.

Procedimiento de análisis

El análisis siguió cuatro pasos. Primero, se extrajeron de la teoría de la validez las inferencias que componen un argumento interpretativo para evaluaciones de desempeño (Kane et al., 1999; Kane, 2013) y los criterios de validación para evaluaciones complejas (Linn et al., 1991; Messick, 1995). Segundo, se identificaron en la literatura sobre aprendizaje inmersivo los rasgos del medio que pueden modificar el desempeño observado o su registro, y se clasificaron según las dos amenazas generales descritas por Messick (1995): la varianza irrelevante al constructo y la subrepresentación del constructo. Tercero, se asignó cada amenaza a la inferencia del argumento que pone en riesgo, siguiendo la lógica aplicada por Cook et al. (2015) a la

simulación clínica. Cuarto, para cada inferencia se derivaron las evidencias y los procedimientos que permitirían comprobarla, con atención a su viabilidad para equipos docentes universitarios. Los resultados se organizan en tablas que sintetizan el argumento, las amenazas y el plan de evidencias.

Alcance y límites

La propuesta es teórica y debe entenderse como una hipótesis de trabajo para la validación, no como evidencia de validez de una evaluación concreta. La selección intencional de las fuentes permite construir un argumento coherente, pero no garantiza la exhaustividad propia de una revisión sistemática. La literatura empírica sobre evaluación en entornos inmersivos procede en su mayor parte de contextos anglosajones, europeos y asiáticos, y de áreas como la salud y la ingeniería, por lo que su transferencia a la docencia universitaria ecuatoriana requiere comprobación.

Resultados

Rasgos de la evaluación en entornos inmersivos que afectan a la validez

El análisis permitió identificar cinco rasgos que distinguen la evaluación en entornos inmersivos de la evaluación convencional y que inciden directamente en la validez. El primero es la mediación del desempeño por una interfaz. En una evaluación escrita u oral el estudiante actúa con recursos que domina desde hace años; en un entorno inmersivo debe, además, desplazarse, seleccionar objetos y comunicarse mediante controles, gestos o un avatar. La revisión de Jensen y Konradsen (2018) mostró que la utilidad de los visores depende de la familiaridad con la tecnología y de la aparición de malestar físico, y el cuestionario de Kennedy et al. (1993) documenta síntomas como náuseas, desorientación y fatiga visual que pueden reducir el desempeño con independencia del aprendizaje.

El segundo rasgo es la carga cognitiva que impone el propio medio. Desde la teoría de la carga cognitiva (Sweller, 1988), la atención que el estudiante dedica a manejar la interfaz o a procesar estímulos irrelevantes compite con la que necesita para resolver la tarea. Makransky et al. (2019) comprobaron que una simulación de laboratorio en realidad virtual inmersiva aumentó la presencia, pero produjo menos aprendizaje que la versión de escritorio, y registraron mayor carga cognitiva en la condición inmersiva. El modelo afectivo-cognitivo del aprendizaje inmersivo (Makransky y Petersen, 2021) explica que la presencia y la agencia pueden favorecer el interés y la autoeficacia, pero también generar sobrecarga. Para la evaluación, esto significa que parte de la varianza de las puntuaciones puede deberse al medio y no a la competencia.

El tercer rasgo es la mayor autenticidad potencial de la tarea. Los entornos inmersivos permiten recrear contextos, roles y consecuencias cercanos a la práctica profesional, lo que coincide con las dimensiones de la evaluación auténtica: tarea, contexto físico, contexto social, resultado y criterios (Gulikers et al., 2004), y con las condiciones de realismo, desafío cognitivo

y juicio evaluativo propuestas por Villarroel et al. (2018). Linn et al. (1991) advirtieron, sin embargo, que la autenticidad aparente de una evaluación de desempeño no sustituye a la evidencia de validez, y que estas evaluaciones presentan problemas de generalización por el número reducido de tareas que pueden aplicarse.

El cuarto rasgo es la disponibilidad de registros detallados del proceso. Las plataformas inmersivas pueden registrar trayectorias, tiempos, secuencias de acciones e interacciones, lo que abre la posibilidad de una evaluación basada en evidencias del proceso y no solo del producto, como propone el diseño centrado en la evidencia (Mislevy et al., 2003) y la evaluación invisible (Shute, 2011). Al mismo tiempo, esta disponibilidad plantea la cuestión de qué registros constituyen evidencia legítima del constructo y cuáles son ruido, y abre problemas de privacidad que la literatura ética ha señalado (O’Brocháin et al., 2016; Slater et al., 2020).

El quinto rasgo es la mediación de la presencia social por avatares. En los mundos virtuales el docente observa una representación del estudiante, no al estudiante. Yee y Bailenson (2007) demostraron que las características del avatar modifican la conducta de quien lo usa, fenómeno conocido como efecto Proteus, y la pérdida o distorsión de las señales no verbales, documentada en la videoconferencia por Bailenson (2021), puede alterar tanto la actuación del estudiante como la percepción del evaluador. Este rasgo introduce una fuente de sesgo que no existe en la observación directa.

Argumento interpretativo y de uso propuesto

A partir de los rasgos descritos se formuló un argumento interpretativo y de uso que encadena cuatro inferencias, siguiendo la estructura propuesta para evaluaciones de desempeño por Kane et al. (1999) y adaptada a la simulación por Cook et al. (2015). La interpretación que se pretende sostener es que la puntuación obtenida por un estudiante en una tarea inmersiva refleja su nivel en la competencia definida en el resultado de aprendizaje y que puede usarse para decidir si la ha alcanzado. La Tabla 2 presenta cada inferencia con su supuesto central y las evidencias que la respaldarían.

Tabla 2. Argumento interpretativo y de uso para la evaluación del aprendizaje en entornos inmersivos

Inferencia	Supuesto que debe sostenerse	Evidencias que lo respaldan
Puntuación	El desempeño observado en el entorno se registra y se califica de forma precisa, consistente y conforme a criterios explícitos, sin que la interfaz o el avatar alteren la observación.	Rúbricas con descriptores observables; acuerdo entre evaluadores; análisis de la severidad de cada evaluador; prueba piloto del registro de la plataforma.
Generalización	La puntuación obtenida en los escenarios, ocasiones y evaluadores concretos es representativa de la que se obtendría en el universo de escenarios y evaluadores admisibles.	Estudio de generalizabilidad con las facetas escenario, ocasión y evaluador; número suficiente de tareas; coeficientes de fiabilidad acordes con el uso.
Extrapolación	La puntuación en el entorno inmersivo refleja la competencia en el dominio real de la práctica profesional y no habilidades ajenas, como el manejo de la tecnología.	Correlación con el desempeño en prácticas reales; análisis de la varianza atribuible a la familiaridad tecnológica, la carga cognitiva y el malestar; juicio de expertos sobre la fidelidad de la tarea.
Decisión	El punto de corte y el uso de la puntuación (aprobación, promoción, retroalimentación) son adecuados, y las consecuencias para los distintos grupos de estudiantes son aceptables.	Método documentado para fijar el punto de corte; análisis de la equidad entre grupos y modalidades; seguimiento de las consecuencias previstas y no previstas.

Nota. Elaboración propia a partir de Kane et al. (1999), Kane (2013), Cook et al. (2015) y Messick (1995).

La inferencia de puntuación exige que el desempeño se observe y se califique con criterios explícitos. En entornos inmersivos esto implica decidir de antemano qué acciones constituyen evidencia de la competencia, cómo se registran y quién las califica. El diseño centrado en la evidencia (Mislevy et al., 2003) ofrece una guía para enlazar el modelo de competencia, el modelo de evidencias y el modelo de tarea, y evita que se califique lo que la plataforma registra con facilidad en lugar de lo que la competencia requiere. La calificación por parte de docentes requiere además criterios compartidos; Sadler (1989) señaló que el juicio cualitativo experto depende de que los evaluadores dispongan de una noción común de calidad, lo que en estos entornos se ve dificultado por la mediación del avatar.

La inferencia de generalización es especialmente frágil en las evaluaciones de desempeño. Linn et al. (1991) y Kane et al. (1999) mostraron que el desempeño varía considerablemente de una tarea a otra, y que una evaluación basada en una o dos tareas complejas difícilmente permite generalizar al dominio. En los entornos inmersivos el costo de diseñar escenarios incentiva el uso de uno solo, lo que agrava el problema. La teoría de la generalizabilidad (Brennan, 2001) permite estimar qué parte de la varianza se debe al estudiante y qué parte a los escenarios, a las ocasiones y a los evaluadores, y calcular cuántos escenarios y evaluadores se necesitan para alcanzar una fiabilidad suficiente para el uso previsto.

La inferencia de extrapolación es la que más se beneficia de la autenticidad del entorno, pero también la que más se ve amenazada por la mediación tecnológica. Su supuesto es que lo que el estudiante hace en el entorno se corresponde con lo que haría en la situación real. La evidencia de que la realidad virtual puede aumentar la presencia sin aumentar el aprendizaje

(Makransky et al., 2019) obliga a comprobar que la puntuación no está dominada por habilidades ajenas al constructo. Cook y Hatala (2016) recomiendan, en simulación, relacionar las puntuaciones con indicadores del desempeño real y con otras medidas del mismo constructo, y examinar las diferencias entre grupos de distinta experiencia.

La inferencia de decisión traslada el análisis a las consecuencias. Messick (1995) incluyó las consecuencias sociales de la evaluación entre los aspectos de la validez, y Moss (1994) argumentó que las evaluaciones complejas requieren un juicio interpretativo que integre distintas fuentes de evidencia. Cuando la calificación inmersiva se usa para aprobar una asignatura o titular a un estudiante de posgrado, el argumento debe demostrar que el punto de corte es defendible y que los estudiantes con menor acceso a la tecnología no resultan sistemáticamente perjudicados.

Amenazas a la validez específicas del medio inmersivo

Las amenazas identificadas se clasificaron según las dos categorías de Messick (1995) y se asignaron a la inferencia que comprometen. Downing y Haladyna (2004) describieron la varianza irrelevante al constructo como la introducción de factores ajenos que aumentan o reducen sistemáticamente las puntuaciones, y la subrepresentación del constructo como la evaluación de solo una parte de lo que se pretende medir. La Tabla 3 sintetiza las amenazas propias del medio inmersivo y las medidas de control que la literatura sugiere.

Tabla 3. Amenazas a la validez propias de la evaluación en entornos inmersivos y medidas de control

Amenaza	Inferencia afectada	Mecanismo	Medida de control
Familiaridad tecnológica desigual	Extrapolación y decisión	Los estudiantes con experiencia previa en videojuegos o visores obtienen ventaja ajena a la competencia.	Sesión de familiarización previa; registro de la experiencia previa; análisis de su relación con las puntuaciones.
Malestar por simulador	Puntuación y extrapolación	Náuseas, desorientación o fatiga visual reducen el desempeño.	Aplicación del cuestionario de malestar por simulador; límites de duración; alternativa de escritorio.
Carga cognitiva extrínseca	Extrapolación	El manejo de la interfaz consume recursos que la tarea necesita.	Interfaces simples; práctica previa; medición de la carga y control estadístico.
Sesgo del evaluador ante el avatar	Puntuación	La apariencia del avatar o la pérdida de señales no verbales influye en la calificación.	Avatares estandarizados; calibración de evaluadores; análisis de severidad con modelos de facetas múltiples.
Fallos técnicos y conectividad	Puntuación y decisión	Interrupciones o latencia alteran el desempeño y el registro.	Protocolo de incidencias; posibilidad de repetición; registro de las condiciones técnicas.
Escenario único	Generalización	Un solo escenario no representa el dominio de situaciones profesionales.	Varios escenarios breves; estudio de generalizabilidad para fijar el número necesario.
Fidelidad visual sin fidelidad de la tarea	Extrapolación	El entorno es realista, pero la tarea no exige la competencia que se quiere evaluar.	Análisis de la tarea con expertos de la profesión; alineación con el resultado de aprendizaje.
Registro del proceso sin modelo de evidencias	Puntuación	Se califican indicadores fáciles de registrar en lugar de los relevantes para el constructo.	Diseño centrado en la evidencia; justificación teórica de cada indicador.
Acceso desigual al equipamiento	Decisión	Los estudiantes con peores condiciones de acceso o con discapacidad quedan en desventaja.	Diseño universal; alternativas equivalentes; análisis de equidad entre grupos y modalidades.

Nota. Elaboración propia a partir de Messick (1995), Downing y Haladyna (2004), Kennedy et al. (1993), Sweller (1988), Makransky et al. (2019), Yee y Bailenson (2007), Linacre (1989), Brennan (2001), Mislevy et al. (2003), Raes et al. (2020) y CAST (2018).

La amenaza de mayor alcance es la varianza irrelevante asociada a la familiaridad tecnológica, porque afecta a la extrapolación y a la decisión al mismo tiempo. Si dos estudiantes con la misma competencia obtienen puntuaciones distintas porque uno de ellos ya ha usado visores, la calificación deja de representar el constructo y genera un problema de equidad. Algo similar ocurre con el malestar por simulador, cuya incidencia varía entre personas (Kennedy et al., 1993), y con la carga cognitiva extrínseca (Sweller, 1988). Estas amenazas no invalidan el uso de entornos inmersivos, pero obligan a medirlas y a controlarlas.

El sesgo del evaluador ante el avatar compromete la inferencia de puntuación. Si las características del avatar modifican la conducta del usuario (Yee y Bailenson, 2007), es plausible que también influyan en la percepción del evaluador, aunque esta hipótesis requiere comprobación empírica. Los modelos de Rasch de facetas múltiples (Linacre, 1989) permiten

estimar la severidad de cada evaluador y ajustar las puntuaciones, y la calibración previa con ejemplos de desempeño anclados contribuye a reducir la variabilidad.

La subrepresentación del constructo aparece cuando la evaluación se limita a un escenario o cuando la fidelidad visual del entorno oculta que la tarea no exige la competencia buscada. Gulikers et al. (2004) subrayaron que la autenticidad depende de la tarea y de los criterios, no solo del contexto físico, y Chernikova et al. (2020) encontraron que los efectos de la simulación dependen del andamiaje y del diseño de la tarea más que del grado de realismo. Por último, la desigualdad de acceso amenaza la inferencia de decisión. La revisión de Raes et al. (2020) sobre el aprendizaje híbrido sincrónico documentó que las condiciones de participación difieren entre estudiantes presenciales y remotos, y el diseño universal para el aprendizaje (CAST, 2018) exige ofrecer medios alternativos de acción y expresión.

Plan de evidencias para la validación institucional

La aplicación del argumento requiere un plan de validación que priorice las inferencias más débiles y las decisiones de mayor impacto. Kane (2013) recomienda que el esfuerzo de validación sea proporcional a la ambición de la interpretación y a las consecuencias del uso: una evaluación formativa de bajo impacto requiere menos evidencia que una evaluación que decide la aprobación o la titulación. La Tabla 4 propone un plan en cuatro fases, viable para equipos docentes universitarios.

Tabla 4. Plan de validación en cuatro fases para evaluaciones en entornos inmersivos

Fase	Acciones	Productos y evidencias
1. Diseño	Definir el constructo y el resultado de aprendizaje; elaborar el modelo de evidencias y la rúbrica; validar el contenido con expertos de la disciplina y de la profesión.	Argumento interpretativo y de uso explícito; rúbrica con descriptores observables; índices de validez de contenido.
2. Piloto	Aplicar la evaluación a un grupo reducido; registrar la experiencia tecnológica previa, el malestar por simulador, la carga cognitiva y las incidencias técnicas.	Ajustes de la tarea y de la interfaz; primeras estimaciones de la varianza irrelevante al constructo.
3. Fiabilidad y generalización	Calificar con al menos dos evaluadores calibrados y varios escenarios; realizar un estudio de generalizabilidad y un análisis de la severidad de los evaluadores.	Coefficientes de generalizabilidad; número necesario de escenarios y evaluadores; informe de sesgos del evaluador.
4. Extrapolación y consecuencias	Relacionar las puntuaciones con el desempeño en prácticas o con otras medidas del constructo; analizar las diferencias por modalidad, acceso y género; consultar al estudiantado.	Evidencias de relación con otras variables; análisis de equidad; decisión documentada sobre el uso de las puntuaciones.

Nota. Elaboración propia a partir de Kane (2013), Cook y Hatala (2016), Brennan (2001) y American Educational Research Association et al. (2014).

La primera fase traslada al diseño el principio de que la validación comienza antes de la aplicación. Explicitar el argumento obliga al equipo docente a decidir qué interpretación quiere sostener y qué evidencias necesitará. La segunda fase permite detectar a tiempo las fuentes de

varianza irrelevante, y la medición de la presencia (Witmer y Singer, 1998), del malestar (Kennedy et al., 1993) y de la carga cognitiva ofrece información para ajustar el entorno. La tercera fase responde a la debilidad conocida de la generalización en las evaluaciones de desempeño, y la cuarta conecta la evaluación con la práctica profesional y con la equidad.

Discusión

El argumento propuesto muestra que la evaluación en entornos inmersivos desplaza el peso de la validación. En la evaluación escrita tradicional la extrapolación suele ser el eslabón más débil, porque la tarea está alejada de la práctica profesional. En los entornos inmersivos la extrapolación puede fortalecerse por la autenticidad del escenario, pero las inferencias de puntuación y de generalización se vuelven más frágiles: la mediación de la interfaz y del avatar introduce nuevas fuentes de error en el registro y la calificación, y el costo de los escenarios reduce el número de tareas. Este desplazamiento coincide con lo que Linn et al. (1991) anticiparon para las evaluaciones de desempeño complejas y con lo que Cook et al. (2015) observaron en la simulación clínica.

Un hallazgo central es que la autenticidad del entorno no garantiza la validez de la calificación. La literatura sobre aprendizaje inmersivo ha tendido a asumir que un escenario más realista produce una evaluación más válida. Sin embargo, la evidencia de que la inmersión puede elevar la presencia y reducir el aprendizaje (Makransky et al., 2019) y la constatación de que los efectos de la simulación dependen más del diseño de la tarea que del realismo (Chernikova et al., 2020) indican que la fidelidad visual puede coexistir con una baja fidelidad de la tarea. Esta distinción tiene implicaciones prácticas: antes de invertir en entornos de alta fidelidad, los equipos docentes deberían asegurarse de que la tarea exige la competencia que quieren evaluar.

La propuesta también pone en primer plano la equidad como componente de la validez. Siguiendo a Messick (1995), las consecuencias de la evaluación forman parte del juicio de validez, y en un país con desigualdades de conectividad y de acceso a equipos, como Ecuador, las diferencias de familiaridad tecnológica pueden convertirse en diferencias de calificación. El argumento obliga a analizar si las puntuaciones funcionan de la misma manera para estudiantes con distintas condiciones de acceso y a ofrecer alternativas equivalentes cuando no sea así.

Implicaciones para la docencia universitaria

Para el profesorado, el argumento ofrece una herramienta de diseño y no solo de verificación. Formular las cuatro inferencias antes de crear la evaluación ayuda a decidir cuántos escenarios usar, qué registrar, cómo calificar y para qué decisiones son adecuadas las puntuaciones. En la práctica, esto se traduce en recomendaciones concretas: incluir una sesión de familiarización antes de evaluar; utilizar varios escenarios breves en lugar de uno largo; calibrar a los evaluadores con ejemplos anclados; registrar las incidencias técnicas y el malestar; y reservar las decisiones de alto impacto para evaluaciones que hayan acumulado evidencias

suficientes. La retroalimentación formativa, en cambio, puede apoyarse en evaluaciones inmersivas con un nivel de evidencia menor, siempre que el docente sea consciente de sus límites.

La propuesta tiene también implicaciones para la formación del profesorado. La validación exige competencias de medición que no suelen formar parte de la preparación docente universitaria, como el análisis de la fiabilidad entre evaluadores o el diseño de rúbricas con descriptores observables. Las unidades de apoyo pedagógico de las universidades podrían asumir esta tarea mediante programas de formación y acompañamiento a los equipos que deseen usar entornos inmersivos en la evaluación.

Implicaciones para la normativa institucional

En el contexto ecuatoriano, la autonomía de las instituciones para regular la evaluación (Consejo de Educación Superior, 2022) permite incorporar exigencias de validación en los reglamentos internos. Las universidades que utilicen entornos inmersivos en evaluaciones de alto impacto, como exámenes de fin de asignatura o defensas de titulación en posgrados en línea, deberían exigir que el diseño incluya un argumento de validez explícito y evidencias proporcionales al uso. Asimismo, el registro de datos de comportamiento en estos entornos debe ajustarse a la Ley Orgánica de Protección de Datos Personales (Asamblea Nacional del Ecuador, 2021), que obliga a informar sobre los fines del tratamiento y a limitar los datos a los necesarios. La literatura ética sobre realidad virtual advierte de los riesgos para la privacidad y la autonomía de estos registros (O’Brolcháin et al., 2016; Slater et al., 2020), lo que refuerza la necesidad de usar solo los indicadores que el modelo de evidencias justifique.

Limitaciones y líneas futuras

La principal limitación es el carácter teórico de la propuesta. El argumento identifica supuestos y amenazas plausibles, pero su comprobación requiere estudios empíricos. La selección intencional de fuentes privilegió los textos fundacionales y las síntesis de investigación, por lo que pueden existir estudios empíricos recientes no considerados. Además, buena parte de la evidencia sobre aprendizaje inmersivo procede de las ciencias de la salud y de la ingeniería, y su transferencia a otras áreas y al contexto latinoamericano debe verificarse. Como líneas futuras se proponen: estudios de generalizabilidad que estimen el número de escenarios y evaluadores necesarios en distintas disciplinas; estudios experimentales sobre el efecto de la apariencia del avatar en el juicio del evaluador; análisis de la invarianza de medida entre estudiantes con distinta familiaridad tecnológica; y estudios de validez de criterio que relacionen las puntuaciones inmersivas con el desempeño en prácticas preprofesionales.

Conclusiones

Este artículo propuso un argumento de validez para la evaluación del aprendizaje en entornos inmersivos en la docencia universitaria, compuesto por cuatro inferencias: puntuación, generalización, extrapolación y decisión. Cada inferencia se acompaña de los supuestos que deben sostenerse, de las amenazas propias del medio y de las evidencias necesarias para respaldarla.

Los entornos inmersivos fortalecen potencialmente la extrapolación al acercar la tarea a la práctica profesional, pero debilitan la puntuación y la generalización. Las amenazas principales son la varianza irrelevante al constructo asociada a la familiaridad tecnológica, al malestar por simulador y a la carga cognitiva extrínseca; la variabilidad entre escenarios, ocasiones y evaluadores; los sesgos del evaluador ante desempeños mediados por avatares; y la desigualdad de acceso.

La autenticidad del escenario no garantiza la validez de la calificación. Por ello, antes de usar las puntuaciones obtenidas en estos entornos para decisiones de alto impacto, los equipos docentes deben explicitar el argumento y reunir evidencias proporcionales al uso. El plan de validación en cuatro fases propuesto (diseño, piloto, fiabilidad y generalización, y extrapolación y consecuencias) ofrece una ruta viable para hacerlo.

Finalmente, la equidad forma parte de la validez. En contextos con desigualdades de acceso tecnológico, como el ecuatoriano, las instituciones deben asegurar que las diferencias en las calificaciones reflejen diferencias de aprendizaje y no de condiciones de acceso, y ofrecer alternativas equivalentes cuando sea necesario.

Referencias

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Asamblea Nacional del Ecuador. (2021). *Ley Orgánica de Protección de Datos Personales*. Registro Oficial Quinto Suplemento n.º 459, 26 de mayo de 2021.
- Bailenson, J. N. (2021). Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior*, 2(1). <https://doi.org/10.1037/tmb0000030>
- Brennan, R. L. (2001). *Generalizability theory*. Springer. <https://doi.org/10.1007/978-1-4757-3456-0>
- CAST. (2018). *Universal Design for Learning guidelines version 2.2*. <http://udlguidelines.cast.org>
- Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research*, 90(4), 499–541. <https://doi.org/10.3102/0034654320933544>
- Consejo de Educación Superior. (2022). *Reglamento de Régimen Académico*. CES.
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, 49(6), 560–575. <https://doi.org/10.1111/medu.12678>

- Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. *Advances in Simulation*, 1, Artículo 31. <https://doi.org/10.1186/s41077-016-0033-y>
- Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830–837. <https://doi.org/10.1046/j.1365-2923.2003.01594.x>
- Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327–333. <https://doi.org/10.1046/j.1365-2923.2004.01777.x>
- Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review*, 27, 244–260. <https://doi.org/10.1016/j.edurev.2019.04.001>
- Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67–86. <https://doi.org/10.1007/BF02504676>
- Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, 8(1), 1–32. <https://doi.org/10.1007/s40692-020-00169-2>
- Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, 23(4), 1515–1529. <https://doi.org/10.1007/s10639-017-9676-0>
- Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational measurement* (4.^a ed., pp. 17–64). American Council on Education; Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17. <https://doi.org/10.1111/j.1745-3992.1999.tb00010.x>
- Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21. <https://doi.org/10.3102/0013189X020008015>
- Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, 33, 937–958. <https://doi.org/10.1007/s10648-020-09586-2>
- Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, 60, 225–236. <https://doi.org/10.1016/j.learninstruc.2017.12.007>
- Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education*, 70, 29–40. <https://doi.org/10.1016/j.compedu.2013.07.033>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, E77-D(12), 1321–1329.

- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS Research Report RR-03-16). Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2003.tb01908.x>
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5–12. <https://doi.org/10.3102/0013189X023002005>
- Mystakidis, S. (2022). Metaverse. *Encyclopedia*, 2(1), 486–497. <https://doi.org/10.3390/encyclopedia2010031>
- O’Brolcháin, F., Jacquemard, T., Monaghan, D., O’Connor, N., Novitzky, P., & Gordijn, B. (2016). The convergence of virtual reality and social networks: Threats to privacy and autonomy. *Science and Engineering Ethics*, 22(1), 1–29. <https://doi.org/10.1007/s11948-014-9621-1>
- Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, 147, Artículo 103778. <https://doi.org/10.1016/j.compedu.2019.103778>
- Raes, A., Detienne, L., Windey, I., & Depaepe, F. (2020). A systematic literature review on synchronous hybrid learning: Gaps identified. *Learning Environments Research*, 23, 269–290. <https://doi.org/10.1007/s10984-019-09303-z>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. <https://doi.org/10.1007/BF00117714>
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. En S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Information Age Publishing.
- Slater, M., Gonzalez-Lienres, C., Haggard, P., Vinkers, C., Gregory-Clarke, R., Jelley, S., Watson, Z., Breen, G., Schwarz, R., Steptoe, W., Szostak, D., Halan, S., Fox, D., & Silver, J. (2020). The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality*, 1, Artículo 1. <https://doi.org/10.3389/frvir.2020.00001>
- Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3, Artículo 74. <https://doi.org/10.3389/frobt.2016.00074>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, 43(5), 840–854. <https://doi.org/10.1080/02602938.2017.1412396>
- Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3), 225–240. <https://doi.org/10.1162/105474698565686>
- Wu, B., Yu, X., & Gu, X. (2020). Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *British Journal of Educational Technology*, 51(6), 1991–2005. <https://doi.org/10.1111/bjet.13023>
- Yee, N., & Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research*, 33(3), 271–290. <https://doi.org/10.1111/j.1468-2958.2007.00299.x>