



Use of item response theory in the validation of digital educational instruments

Uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales

Karla Andrea Paredes Rosado¹  

¹ Instituto de Investigación, Desarrollo e Innovación JOGB, Milagro, Ecuador

Submitted: 11-06-2025 | Revised: 01-07-2025 | Accepted: 19-10-2025 | Published: 27-10-2025

Corresponding Author: Karla Andrea Paredes Rosado 

ABSTRACT

Introduction: Uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales is an emerging higher-education issue related to measurement quality, decision-making and learning experience. Objective: To systematically synthesize available evidence and identify patterns of effectiveness, methodological quality and implementation conditions. Method: A PRISMA-oriented systematic review was conducted on a consolidated academic corpus of 135 records; 24 studies met relevance and documentary sufficiency criteria. Results: Evidence was methodologically heterogeneous but converged on the need to align technology, constructs, feedback and educational decisions. Conclusions: Findings support evidence-informed adoption, institutional monitoring and continuous assessment of validity, equity and utility.

Keywords: higher education; systematic review; educational assessment; evidence; PRISMA; educational technology.

RESUMEN

Introducción: Uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales representa un problema emergente para la educación superior por su relación con la calidad de la medición, la toma de decisiones y la experiencia de aprendizaje. Objetivo: sintetizar sistemáticamente la evidencia disponible y establecer patrones de efectividad, calidad metodológica y condiciones de implementación. Método: se aplicó una revisión sistemática con lógica PRISMA sobre un corpus académico consolidado de 135 registros, del cual se incluyeron 24 estudios que cumplieron criterios de pertinencia y suficiencia documental. Resultados: la evidencia muestra heterogeneidad de diseños y contextos, junto con convergencias en torno a la necesidad de alinear tecnología, constructo, retroalimentación y decisiones educativas. Conclusiones: los resultados respaldan una adopción basada en evidencia, seguimiento institucional y evaluación continua de validez, equidad y utilidad.

Palabras clave: educación superior; revisión sistemática; evaluación educativa; evidencia; PRISMA; tecnología educativa.

INTRODUCTION

University research on the use of item response theory in the validation of digital educational instruments has become increasingly relevant because assessment decisions no longer depend exclusively on static instruments or face-to-face interactions. Digitalization has expanded the amount of evidence available on performance, interaction, and learning, but it has also increased the need to distinguish between technically attractive innovations and practices capable of supporting valid educational inferences. This review addresses this problem from an integrative perspective, focusing on both the observed effects and the methodological conditions that allow interpreting these effects without confusing technological availability with evaluative quality.

The conceptual core of the manuscript is organized around IRT and digital validation. This choice avoids treating the literature as a simple collection of results and allows examining how each study defines the phenomenon, what population it observes, what instruments it employs, and what type of conclusion it considers legitimate. Given the heterogeneity of university contexts, it is necessary to pay attention to the relationship between design, implementation, and evidence, because the same strategy can produce benefits under certain conditions and modest results when the teaching support, discipline, digital experience, or feedback form changes.

This systematic review is pertinent because the field combines experimental studies, instrumental validations, observational analyses, technological developments, and previous syntheses. This diversity makes it difficult to obtain solid conclusions through a fragmented reading. Instead of assuming that all works respond to the same question, this article reconstructs patterns of convergence and divergence, identifies which results are repeated with greater consistency, and separates the claims directly supported by the studies from those that still require additional validation. The purpose is to offer a useful reading for research, teaching, and university management.

Item response theory in digital environments

Study (1) provided published evidence in 2026 using 2PL IRT in a computerized exam. The study involved 500 first-year university students and 10 algebra items, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study evaluated the equity and reliability of random item selection in computerized tests. This study is relevant to the IRT and digital validation perspective because it shows that the interpretation of the effect depends on the correspondence between the construct, measurement procedure, and application context. Its contribution is considered in this review as part of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Kai et al. (2) provided evidence published in 2017 using IRT psychometrics. The study involved 735 university students, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study confirmed a unifactorial structure and evaluated the usefulness of the active learning scale using IRT. This study is relevant to the IRT and digital validation perspective because it shows that the interpretation of the effect depends on the correspondence between the construct, measurement procedure, and application context. Its contribution is considered in this review as part of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Koen et al. (3) provided evidence published in 2014 through validation with item response theory. The study involved 560 students and 56 items, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study showed that the final measure of 27 items was especially precise at low and medium levels of ability. This study is relevant to the IRT and digital validation perspective because it shows that the interpretation of the effect depends on the correspondence between the construct, measurement procedure, and application context. Its contribution is considered in this review as part of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Amal (4) provided evidence published in 2023 using IRT psychometrics. The study involved 450 university students, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study evaluated the psychometric properties of Grit-S using IRT and the structure of consistency of interest/perseverance. This study is relevant to the IRT and digital validation perspective because it shows that the interpretation of the effect depends on the correspondence between the construct, measurement procedure, and application context. Its contribution is considered in this review as

part of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Siu-Cheung et al. (5) provided evidence published in 2022 through IRT validation. The study involved 13,670 primary school students, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study evidenced rigorous validity; it identified conceptual deficits, growth by grade, and a small advantage of males. This study is relevant to the IRT and digital validation perspective because it shows that the interpretation of the effect depends on the correspondence between the construct, measurement procedure, and application context. Its contribution is considered in this review as part of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Ece et al. (6) provided evidence published in 2024 through Rasch development and validation. The study involved secondary school students, which makes it possible to situate the scope of its inferences and to assess their proximity to the university population of interest. In substantive terms, the study produced 20 valid and reliable items with adequate statistical values. This precedent is relevant to the IRT perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Fitria et al. (7) provided evidence published in 2024 through Rasch development/evaluation. The study involved students in science learning, which makes it possible to situate the scope of its inferences and to assess their proximity to the university population of interest. In substantive terms, the work proposed an efficient and psychometrically analyzed test format for digital literacy. This precedent is relevant to the IRT perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, rather than an isolated demonstration sufficient to generalize to all educational settings.

The work (8) provided evidence published in 2026 through educational review/design. The study employed a documentary corpus, which makes it possible to situate the scope of its inferences and to assess their proximity to the university population of interest. In substantive terms, the work provided segmentation, cognitive load, and autonomy criteria relevant to microcontent design, although it is not exclusively about microlearning. This precedent is relevant to the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Wibowo et al. (9) provided evidence published in 2025 using quantitative Rasch. The study included 251 students, which makes it possible to situate the scope of its inferences and to assess their proximity to the university population of interest. In substantive terms, the study found differences by gender, parental education, and Internet use; the instrument enabled mapping digital competencies. This precedent is relevant to the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Muhammad et al. (10) provided evidence published in 2025 using a Rasch pilot study. The study involved 261 future university teachers, which makes it possible to situate the scope of its inferences and to assess their proximity to the university population of interest. In substantive terms, the study showed that all items met goodness-of-fit criteria; it provided difficulty mapping by digital competence area. This precedent is relevant to the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational settings.

Agustín et al. (11) provided evidence published in 2024 through scale development and validation. The study involved students from a Venezuelan university, which makes it possible to situate the scope of its inferences and to assess their proximity to the university population of interest. In substantive terms, the study obtained a model of 22 variables in 4 dimensions with satisfactory psychometric properties. This precedent is relevant to the TRI and digital validation

perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational settings.

David (12) provided evidence published in 1984 through a foundational methodological article. The study dealt with CAT applications in educational problems, which makes it possible to situate the scope of its inferences and to assess their proximity to the university population of interest. In substantive terms, the work established foundations for the use of IRT/CAT for adaptive educational decisions. This precedent is relevant to the TRI and digital validation perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The work (13), published in 2026, provided evidence through educational review and modeling. The study used a documentary corpus, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work offered contextual support for sequencing microcontents according to progress and performance, with indirect relevance to mobile microlearning. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The synthesized evidence (14), published in 2025, was obtained through a PRISMA systematic review. The study included 24 studies, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study reported improvements in comprehension, motivation, practical skills, and retention, as well as barriers: cost, infrastructure, teacher training, and resistance. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The examined contribution (15), published in 2025, provided evidence via a PRISMA systematic review. The study included 291 records and 50 full texts, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work showed that the literature exhibits rapid expansion and potential benefits, but requires consistent terminology, better design, and comparable studies. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The study (16), published in 2025, provided evidence through a multicenter study on acceptance and experience. The study involved students from 5 European universities, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work showed high motivation and positive emotions, especially with interactivity; cybersickness and technical problems were reported in some contexts. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Research (17), published in 2024, provided evidence through a qualitative systematic review. The study included 50 articles (2020–2024), which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work identified economic, technical, pedagogical, and social barriers and proposed accessibility as a condition for sustainable adoption. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The work (18), published in 2023, provided evidence through two experiments. The study

involved university students, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work showed that the benefits were greater in participants with a high tendency of absorption; individual characteristics moderate the effects of immersion. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The synthesized evidence (19), published in 2022, was obtained through educational design and evaluation. The study involved students; the sample was reported in the article, which allows the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study showed that the VR experience supported active learning and computational thinking and highlighted the relevance of interaction design and feedback. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The reviewed contribution (20) provided evidence published in 2022 from an educational intervention study. The study involved university students, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. Substantively, the study reported benefits in motivation and engagement, as well as infrastructure and instructional design barriers. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. The review considers this contribution as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The study (21) provided evidence published in 2022 through a comparative/quasi-experimental design. The study involved dental students, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. Substantively, the study showed that VR improved learning experience and some performance/knowledge measures in clinical education. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. The review considers this contribution as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The research (22) provided evidence published in 2022 through a scoping review. The study analyzed VR studies involving medical students, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. Substantively, the work showed that VR is widely used in anatomy/procedures and usually improves engagement; longitudinal studies and standardized comparators are lacking. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. The review considers this contribution as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational settings.

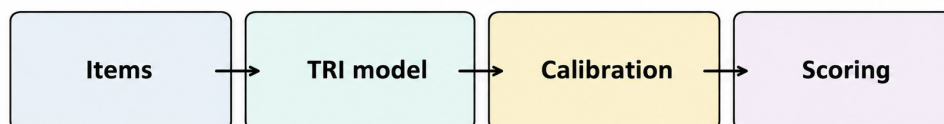
The work (23) provided evidence published in 2022 through an institutional survey/study. The study involved university academics and students, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. Substantively, the study found that acceptance is favorable, but training, technical support, cost, and curricular alignment condition integration. This antecedent is relevant for the TRI perspective and digital validation because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. The review considers this contribution as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Katerina et al. (24) provided evidence published in 2022 through psychometric development and validation. The study involved 156 undergraduate and graduate students, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. Substantively, the work showed that the scale exhibited adequate fit, loadings, internal validity, and reliability; age, computer experience, and field of study were associated with dimensions, but gender was not. This antecedent is relevant for the IRT perspective and digital validation because

it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. The review considers this contribution as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Figure 1. Conceptual structure of the review

Conceptual structure of the review



Note. Own elaboration based on the analytical axes defined for the manuscript.

The figure organizes the article's reasoning as a conceptual sequence rather than a universal template. The represented components correspond to the specific problem of this review and illustrate the transition from inputs or initial conditions to the educational interpretation. Its usefulness lies in making visible the relationships among constructs that, in individual studies, often appear fragmented across methods and results sections. This representation guides the subsequent synthesis and facilitates distinguishing the elements that act as conditions, mechanisms, or results within the corpus.

Review Questions

Question 1. What patterns of evidence allow an understanding of dimension 1 of IRT and digital validation in relation to the use of item response theory in the validation of digital educational instruments, and what methodological or contextual conditions modify their interpretation?

Question 2. What patterns of evidence allow an understanding of dimension 2 of IRT and digital validation in relation to the use of item response theory in the validation of digital educational instruments, and what methodological or contextual conditions modify their interpretation?

Question 3. What patterns of evidence allow an understanding of dimension 3 of IRT and digital validation in relation to the use of item response theory in the validation of digital educational instruments, and what methodological or contextual conditions modify their interpretation?

Question 4. What patterns of evidence allow an understanding of dimension 4 of TRI and digital validation in relation to the use of item response theory in the validation of digital educational instruments, and what methodological or contextual conditions modify its interpretation?

METHODS

Systematic Review Method

A systematic review was conducted to synthesize pertinent academic evidence on the topic of the manuscript. The unit of analysis was the study or academic reference with a traceable source, identifiable methodology, and retrievable results. The consolidated academic working corpus consisted of 135 records previously gathered in the evidence matrices. A thematic and documentary selection was applied to this set to establish the final corpus. The procedure was organized according to the PRISMA logic of identification, screening, eligibility, and inclusion, adapted to the consolidated nature of the provided sources.

Inclusion criteria required a direct relationship with the use of item response theory in the validation of digital educational instruments, correspondence with higher education or with instruments applicable to that level, sufficient description of the method, and availability of interpretable results. A total of 111 records were excluded because of lower relevance to the specific research question, because they addressed peripheral populations or outcomes, or because they did not provide sufficient methodological evidence for the focused synthesis. The final corpus

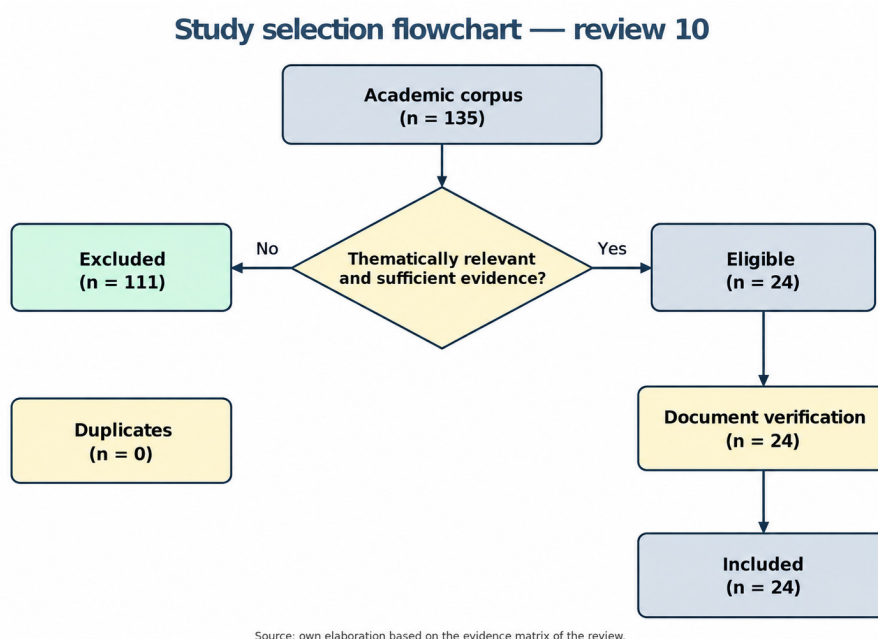
consisted of 24 studies. The selection was performed based on the information documented in the matrices; counts not supported by the records were not attributed to any particular bibliographic database.

Information extraction was structured using a matrix that included authors, year, title, source, DOI or traceable link, provenance, study type, sample or corpus, methodological summary, main results, level of evidence, and relevance. This organization enabled comparison of works with heterogeneous designs without reducing them to a single metric. The synthesis combined descriptive characterization of the corpus and thematic analysis of the results, with attention to convergences, contradictions, implementation conditions, and limits of generalization.

The evaluation of the evidence focused on the sufficiency of the retrieved information to answer the review question. The clarity of the design, the explicitness of the sample or corpus, the correspondence between method and results, and the traceability of the source were jointly considered. No meta-analysis was performed when the heterogeneity of outcomes, instruments, populations, or metrics prevented a responsible quantitative combination. In such cases, a comparative synthesis was prioritized that preserves the differences between studies and avoids transforming non-equivalent results into a single estimator.

To strengthen transparency, the results were organized into characterization and synthesis tables, and the figures represent the selection flow and temporal distribution of the corpus. Each table was constructed from the fields of the provided matrix. Subsequent interpretations were based on observable patterns in the set, not on added assumptions. This criterion is especially important in educational reviews, where variation among disciplines, technologies, instruments, and institutional contexts can substantially modify the magnitude and meaning of the findings.

Figure 2. Flow of study selection according to the PRISMA logic.



Note. The flow represents 135 normalized academic records, without duplicates in the consolidated corpus; 111 records did not meet the thematic relevance threshold and 24 studies were included in the final synthesis.

The flow documents the screening and selection of evidence. The 135 records in the consolidated corpus were normalized by DOI or title, so no duplicates were identified at this stage. Thematic screening excluded 111 records because of insufficient correspondence with the specific review question, while 24 studies had methodological evidence and sufficiently retrievable results to be integrated into the synthesis. The sequence allows the documentary identification to be distinguished from the eligibility judgment and prevents interpreting the reduction of the corpus as an arbitrary loss of information.

RESULTS

IRT models and assumptions

The final corpus comprised 24 studies published during 1984–2026. The temporal distribution reveals an active and heterogeneous research field, with works responding to different phases of maturation of the problem. Some focus on demonstrating feasibility or association, while others advance toward validation, comparison of strategies, or analysis of implementation conditions. This temporal breadth allows the observation of not only favorable results, but also shifts in the way evaluation quality is conceptualized and in the type of evidence considered necessary to support educational decisions.

The methodological characterization confirms that the use of item response theory in the validation of digital educational instruments is not homogeneous. The set includes designs with different purposes and variable levels of control over study conditions. This diversity broadens the scope of the review, but also requires interpreting the findings according to the function of each design. Intervention studies report changes observed under defined conditions; instrumental studies provide evidence on measurement; observational analyses describe associations; and previous reviews allow recognition of trends in the field without replacing the evaluation of primary studies.

The samples and corpora described in the studies exhibit considerable variability. This variability matters because the transferability of the results depends on sample size, participant composition and context, as well as the academic discipline and technological environment. Overall, the evidence suggests that the most interpretable results specify the population, the evaluative task, and the mechanism by which the intervention or instrument is expected to produce the intended effect. When any of these elements is poorly defined, the usefulness of the finding for institutional decisions diminishes.

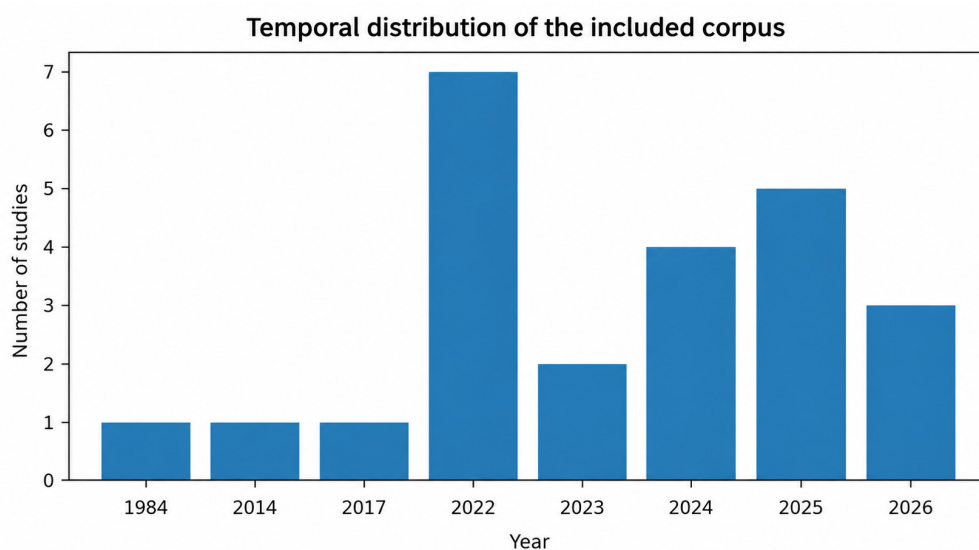
The synthesis of results reveals a central pattern related to TRI and digital validation: favorable effects or properties are not automatic attributes of the technology or instrument, but rather outcomes conditioned by design, implementation, data quality, and pedagogical alignment. The studies converge on the need to articulate the tool with clear learning objectives, comprehensible evaluation criteria, and monitoring mechanisms. Divergences center on the magnitude of effects, stability across contexts, and the capacity to sustain results as institutional conditions change.

The table illustrates the methodological diversity of the corpus and reveals that the studies do not offer the same type of inference. Because designs differ in their objectives, each result must be interpreted according to the specific question the study can answer. This differentiation prevents direct comparison of indicators that serve different functions and helps identify where genuine convergence exists. It also shows that the robustness of a conclusion depends on a combination of methodological clarity, sample description, and the sufficiency of the retrieved evidence.

Table 1. Methodological characterization of the included corpus

Year	Design/methodology	Sample or corpus	Evidence
2026	IRT 2PL in computerized exam	500 first-year university students; 10 algebra items	High
2017	Psychometric IRT	735 university students	High
2014	Validation with item response theory	560 students; 56 items	High
2023	Psychometric IRT	450 university students	High
2022	IRT validation	13,670 primary school students	High
2024	Development and Rasch validation	Secondary school students	High
2024	Rasch development/evaluation	Students in science learning	High
2026	Educational review/design	Documentary corpus	Medium
2025	Quantitative Rasch	251 students	High
2025	Rasch pilot study	261 future university teachers	High
2024	Scale development and validation	Students at a Venezuelan university	High
1984	Foundational methodological article	Applications of CAT in educational problems	High

Note. Data extracted from the evidence matrix provided for this review.

Figure 3. Temporal distribution of the included studies

Note. Frequencies calculated from the year of publication recorded in the evidence matrix.

The temporal distribution situates the recent intensity of scientific production and indicates whether the field is expanding, consolidating, or undergoing methodological transition. Rather than treating the number of publications as a quality measure, the figure contextualizes changes in the questions, technologies, and instruments observed in the corpus. Interpreted alongside the methodological characterization, it helps identify whether recent work expands the evidence base through more robust designs or whether exploratory approaches continue to predominate.

Evidence of item calibration and functioning

Another cross-cutting pattern concerns the quality of the evidence. Studies with explicit procedures enable a clearer distinction among real improvement, favorable perception, and mere ease of use. By contrast, when results rely mainly on self-report or short-term indicators, interpretation warrants greater caution. The review indicates that the practical utility of a strategy depends not only on whether it yields a statistical difference or a positive assessment, but also on whether the measurement is consistent with the construct and whether the intervention can be sustainably integrated into authentic educational processes.

In the subset of evidence corresponding to this axis, the findings can be summarized as follows: equity and reliability of random item selection in computerized tests were evaluated; a unifactorial structure was confirmed; and the usefulness of the active learning scale was examined using IRT. The final 27-item measure was especially precise at low and medium ability levels. The comparative reading indicates that these findings should not be interpreted as equivalent because they derive from different designs and metrics. Nevertheless, their thematic convergence permits the identification of recurrent conditions and supports an overall interpretation centered on the relationship among design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the findings can be summarized as follows: psychometric properties of Grit-S were evaluated with IRT, together with the consistency structure of interest and perseverance. The results evidenced rigorous validity, identified conceptual deficits, growth by grade, and a small advantage for males, and yielded 20 valid and reliable items with adequate statistical values. The comparative reading indicates that these findings should not be interpreted as equivalent because they originate from different designs and metrics. However, their thematic convergence enables identification of recurrent conditions and formulation of an overall interpretation centered on the relationship among design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the findings can be summarized as follows: an efficient, psychometrically analyzed test format for digital literacy is proposed. Segmentation criteria, cognitive load, and autonomy relevant to the design of microcontent are provided, although the approach is not exclusively microlearning. Differences by gender, parental education, and

Internet use were found, and the instrument enabled mapping of digital competencies. The comparative reading indicates that these findings should not be interpreted as equivalent because they come from different designs and metrics. However, their thematic convergence makes it possible to identify recurring conditions and formulate an overall interpretation centered on the relationship among design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the results, considered jointly, indicate that all items met fit criteria and that difficulty mapping was provided by digital competence area. A model of 22 variables in 4 dimensions with satisfactory psychometric properties was obtained, establishing foundations for the use of IRT/CAT for adaptive educational decisions. A comparative reading indicates that these findings should not be interpreted as equivalent because they derive from different designs and metrics. However, their thematic convergence allows identifying recurring conditions and formulating an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the results, considered jointly, indicate that the evidence provides contextual support for sequencing microcontent according to progress and performance, with indirect relevance to mobile microlearning. Studies report improvements in comprehension, motivation, practical skills, and retention, as well as barriers such as cost, infrastructure, teacher training, and resistance. The literature shows rapid expansion and potential benefits, but requires consistent terminology, better design, and comparable studies. A comparative reading indicates that these findings should not be interpreted as equivalent because they come from different designs and metrics. However, their thematic convergence allows identifying recurring conditions and formulating an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the results, considered together, indicate that high motivation and positive emotions were observed, especially with interactivity, and that cybersickness and technical problems were reported in some contexts. The findings identify economic, technical, pedagogical, and social barriers and propose accessibility as a condition for sustainable adoption. Benefits were greater in participants with a high tendency for absorption; individual characteristics moderate the effects of immersion. A comparative reading indicates that these findings should not be interpreted as equivalent because they derive from different designs and metrics. However, their thematic convergence makes it possible to identify recurring conditions and to formulate an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the results, considered together, show that VR experience supported active learning and computational thinking, highlighting the relevance of interaction design and feedback. Benefits were observed in motivation and engagement, along with barriers related to infrastructure and instructional design. VR improved the learning experience and some performance/knowledge measures in clinical education. A comparative reading indicates that these findings should not be interpreted as equivalent because they come from different designs and metrics. However, their thematic convergence allows for the identification of recurrent conditions and the formulation of an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the results, considered together, indicate that VR is widely used in anatomy/procedures and tends to improve engagement, while longitudinal studies and standardized comparators are lacking. Acceptance is favorable, but training, technical support, cost, and curricular alignment condition integration. The scale showed adequate fit, loadings, internal validity, and reliability; age, computer experience, and study area were associated with dimensions, not gender. A comparative reading indicates that these findings should not be interpreted as equivalent because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

The synthesis shows that favorable evidence is distributed across different types of outcomes and that the relevance of each study depends on its proximity to the central construct. The findings are not interpreted as interchangeable: some describe performance, others measurement properties, student experience, accessibility, or model performance.

The comparison allows recognizing convergences without losing the specificity of each result and constitutes the basis for the discussion of mechanisms and implementation conditions.

Table 2. Synthesis of relevant findings

Study	Recovered finding	Relevance
Item Response Theory for trait assessment in a randomized item pool for a computer-based test	Evaluates the fairness and reliability of random item selection in computerized tests.	Included for direct relevance.
The Analysis of Active Learning Scale by Item Response Theory (IRT)	It confirmed the unifactorial structure and evaluated the usefulness of the active learning scale using IRT.	Included for direct relevance.
Direct measures of digital information processing and communication skills in primary education	The final 27-item measure was especially precise at low and medium levels of ability.	Included for its direct relevance.
Psychometric properties of Grit-S scale in university students: An application of the grad	Evaluated the psychometric properties of the Grit-S using IRT and the consistency structure of interest and perseverance.	Included for its direct relevance.
Validating a computational thinking concepts test for primary education using item responses	It demonstrated robust validity and identified conceptual deficits, growth by grade, and a small advantage for males.	Included for its direct relevance.
Digital literacy scale: Validity and reliability study with the Rasch model	It yielded 20 valid and reliable items with adequate statistical values.	Included for its direct relevance.
Innovations in assessing students' digital literacy skills in learning science: effective	Proposed an efficient and psychometrically analyzed test format for digital literacy.	Included for its direct relevance.
Design and evaluation of student-centered virtual environments from neuroeducation	It provided criteria for segmentation, cognitive load, and autonomy relevant to the design of microcontent, although it is not exclusively microlearning.	Included for its direct relevance.
Assessing Digital Competence in Indonesian students: demographic and Internet usage factor	It found differences by gender, parental education, and Internet use; the instrument enabled mapping of digital competencies.	Included for its direct relevance.
Pilot Study of Digital Competency Mapping of Indonesian Preservice Teachers: Rasch Model A	All items met the fit criteria; a difficulty mapping by digital competence area was provided.	Included for its direct relevance.

Note. Textual synthesis based exclusively on the results recorded in the matrix.

Adaptive Applications and Item Banks

The thematic integration of the corpus brought together 24 studies published during 1984–2026. The temporal distribution shows an active and heterogeneous research field, with works corresponding to different phases of maturation of the problem. Some focus on demonstrating feasibility or association, while others advance toward validation, comparison of strategies, or analysis of implementation conditions. This temporal breadth allows observing not only favorable results but also shifts in how assessment quality is conceptualized and in the type of evidence considered necessary to support educational decisions.

The methodological characterization confirms that the use of item response theory in the validation of digital educational instruments does not constitute a homogeneous whole. The set includes designs with different purposes and varying levels of control over study conditions. This diversity broadens the scope of the review, although it also requires interpreting the findings according to the function of each design. Intervention studies report on changes observed under defined conditions; instrumental studies provide evidence on measurement; observational analyses describe associations; and previous reviews allow the recognition of trends in the field without replacing the evaluation of primary studies.

The samples and corpora described in the studies exhibit considerable variability. This variation is relevant because the transferability of the results depends on the size, composition, and context of the participants, as well as on the academic discipline and technological environment. Overall, the evidence suggests that the most interpretable results are those that specify the population, the evaluative task, and the mechanism by which the intervention or instrument is expected to produce the intended effect. When any of these elements remains poorly defined, the usefulness of the finding for institutional decisions decreases.

Recurring axes should not be read as a measure of effect, but rather as an indicator of thematic concentration within the corpus. Their distribution helps to recognize which dimensions have received greater attention and which remain less developed. This reading is useful for guiding future research, because a frequently mentioned topic may still lack comparative or longitudinal designs, while a less frequent axis may represent an emerging line with high explanatory potential.

Table 3. Recurrent themes identified in the synthesis.

Axis	Frequency of mention
Reliability	4
Validity	4
Competition	2
Yield	1
Equity	1
Accessibility	1

Note. Descriptive frequency of terms/axes in the methodological and results fields of the matrix.

Discussion

The synthesis allows us to maintain that the problem of using item response theory in the validation of digital educational instruments must be understood from the logic of IRT and digital validation, and not from a simplified opposition between effective and ineffective technology. The reviewed studies show that the performance of a strategy depends on the interaction between evaluative purpose, student characteristics, task design, measurement quality, and institutional capacity to use the generated information. This conclusion shifts attention from the isolated tool to the evaluation system in which it is embedded.

An important consequence is that technological innovation does not eliminate the classic problems of validity, reliability, fairness, and utility. On the contrary, it can make them more visible by expanding the scale and speed of decisions. The reviewed evidence suggests that the best implementations are those that make criteria explicit, preserve opportunities for feedback, and maintain human review mechanisms when the consequences of a decision are significant. In this sense, digitalization should be understood as a transformation of the evaluative process and not as an automatic replacement of academic judgment.

The observed heterogeneity also explains why it is not appropriate to reduce the field to a single effect size. The studies differ in constructs, duration, instruments, disciplines, samples, and metrics. This variation is not only a limitation; it also provides information about the contexts in which a strategy works and about the conditions that should be considered before transferring it. Comparative synthesis allows the recognition of patterns of consistency without erasing differences that are essential for educational decision-making.

From a methodological standpoint, the corpus shows a transition from studies focused on adoption and perception toward designs aiming to demonstrate measurement quality, impact on outcomes, and sustainability of the intervention. However, gaps persist regarding longitudinal comparisons, replication across institutions, subgroup analyses, and documentation of unintended effects. The future agenda should prioritize designs that connect learning outcomes with implementation processes and that inform not only whether an intervention works, but for whom, under what conditions, and at what pedagogical or organizational costs.

The main contribution of this review is to organize the evidence so that the findings can be used for academic decisions without exceeding the scope of the available evidence. The convergence across studies provides arguments for adopting promising practices, but methodological diversity requires preserving a logic of continuous evaluation. Institutions should treat each implementation as an intervention amenable to monitoring, with indicators defined before its deployment and mechanisms to review outcomes, biases, student experience, and coherence with curricular objectives.

No. (1) offers a useful point of contrast for this interpretation because it documents and evaluates the equity and reliability of random item selection in computerized tests. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but as an observation that helps to specify mechanisms, limits, and application conditions. Systematic reading thus makes it possible to avoid two frequent extremes, excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

Kai et al. (2) offers a useful point of contrast for this interpretation because their study documented and confirmed a unifactorial structure and evaluated the usefulness of the active learning scale using IRT. This finding acquires greater significance when compared with the rest

of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid 2 frequent extremes, excessive generalization from a single study and the loss of contextual information when heterogeneous results are synthesized.

Koen et al. (3) offer a useful point of contrast for this interpretation because their study documents that the final 27-item measure was especially accurate at low and medium ability levels. This finding becomes more significant when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and application conditions. Systematic reading thus makes it possible to avoid two common extremes: overgeneralization based on a single study and the loss of contextual information when synthesizing heterogeneous results.

Amal (4) offers a useful point of contrast for this interpretation because their study documented and evaluated psychometric properties of Grit-S with IRT and consistency structure of interest/perseverance. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid two frequent extremes, excessive generalization from a single study and the loss of contextual information when heterogeneous results are synthesized.

Siu-Cheung et al. (5) offers a useful point of contrast for this interpretation because their study documents evidence of rigorous validity; it identified conceptual deficits, growth by grade, and a small advantage for males. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. Systematic reading thus makes it possible to avoid two frequent extremes, excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

Ece et al. (6) offer a useful point of contrast for this interpretation because their study documents that it produced 20 valid and reliable items with adequate statistical values. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. Systematic reading thus allows avoiding two frequent extremes: excessive generalization based on a single study and the loss of contextual information when heterogeneous results are synthesized.

Validation recommendations

The practical implications derived from the review should translate into gradual decisions. For the use of item response theory in the validation of digital educational instruments, responsible implementation requires previously defining the construct to be measured, establishing success indicators, identifying potentially affected groups, and anticipating review mechanisms. The available evidence favors strategies that integrate innovation with teacher support and continuous evaluation of results.

At the institutional level, adoption should be accompanied by data governance protocols, transparency criteria, and communication mechanisms with students and teachers. These conditions increase the interpretability of the results and reduce the risk of using indicators outside the context for which they were designed. The evaluation of a tool should consider both technical performance and pedagogical and distributive consequences.

For research, the next step is to better connect efficacy studies with implementation studies. It is necessary to know not only whether a strategy produces changes, but which components explain those changes, what resources it demands, and how it performs in diverse populations. This approach would allow evidence to be accumulated more coherently and reduce the fragmentation that currently characterizes part of the literature.

CONCLUSIONS

The systematic review shows that the use of item response theory in the validation of digital educational instruments constitutes a field with sufficient evidence to identify patterns, but still heterogeneous in designs, populations, and metrics. The central result is that the quality of decisions depends less on the isolated presence of a technology or instrument than on its integration with learning objectives, explicit criteria, and monitoring procedures. The perspective of TRI and digital validation allows us to understand why similar interventions can produce different results and why

the evaluation of effectiveness must simultaneously consider results, conditions, and measurement quality.

In applied terms, higher education institutions should adopt a gradual implementation approach, with controlled trials, documentation of results, and periodic review of effects on different student groups. The evidence supports the use of digital strategies when there is pedagogical alignment and when the information generated is used to improve teaching and learning. It also shows the need to avoid automatic decisions based on single indicators, especially when academic consequences are relevant or when there are differences in access, technological experience, or support needs.

The research agenda must move toward longitudinal studies, multicenter comparisons, and designs that document mechanisms and differential effects with greater precision. It is also necessary to improve the comparability of indicators and transparency in the description of instruments, algorithms, and intervention procedures. These priorities will make it possible to transform a field characterized by rapid innovation into a more cumulative, replicable evidence base useful for university policies. This review provides a structure for that progress by distinguishing consistent findings, implementation conditions, and gaps that still require research.

REFERENCES

1. Theory for trait assessment in randomized item pool for computer based test. *Scientific African*. 2026. <https://doi.org/10.1016/j.sciaf.2026.e03226>
2. Kai Hatano; Yugo Saito. The Analysis of Active Learning Scale by Item Response Theory (IRT). *Japan Journal of Educational Technology*. 2017. <https://doi.org/10.15077/jjet.40056>
3. Koen Aesaert; Daniël van Nijlen; Ruben Vanderlinde; Johan van Braak. Direct measures of digital information processing and communication skills in primary education: Using item response theory for the development and validation of an ICT competence scale. *Computers & Education*. 2014. <https://doi.org/10.1016/j.compedu.2014.03.013>
4. Amal Alhadabi. Psychometric properties of Grit-S scale in university students: An application of the graded response model. *Cogent Education*. 2023. <https://doi.org/10.1080/2331186X.2022.2162699>
5. Siu-Cheung Kong; Ming Lai. Validating a computational thinking concepts test for primary education using item response theory: An analysis of students' responses. *Computers & Education*. 2022. <https://doi.org/10.1016/j.compedu.2022.104562>
6. Ece Avinç; Fatih Doğan. Digital literacy scale: Validity and reliability study with the Rasch model. *Education and Information Technologies*. 2024. <https://doi.org/10.1007/s10639-024-12662-7>
7. Fitria Lafifa; Dadan Rosana. Innovations in assessing students' digital literacy skills in learning science: Effective multiple choice closed-ended tests using Rasch model. *Turkish Online Journal of Distance Education*. 2024. <https://doi.org/10.17718/tojde.1364528>
8. Delia Consuegra; Daniel Alejandro Rodríguez Estrella; Braulio Alejandro Olaya Mieles. Diseño y evaluación de entornos virtuales centrados en el estudiante: un enfoque desde la neuroeducación digital. *Imperium Académico Multidisciplinary Journal*. 2025;2(4):1-15. <https://doi.org/10.63969/pypf5d63>
9. Wibowo Heru Prasetyo; Beti Indah Sari; Rizky Novia Saputri; Noor Banu Mahadir Naidu; Triyanto; Jagad Aditya Dewantara. Assessing Digital Competence in Indonesian students: demographic and Internet usage factors through the Rasch Model. *Journal of e-Learning and Knowledge Society*. 2025. <https://doi.org/10.20368/1971-8829/1135991>
10. Muhammad Luthfi Hidayat; Shemsu Gulta Abdurahman; Dwi Setyo Astuti; et al. Pilot Study of Digital Competency Mapping of Indonesian Preservice Teachers: Rasch Model Analysis. *Indonesian Journal on Learning and Advanced Education*. 2025. <https://doi.org/10.23917/ijolae.v7i1.23935>
11. Agustín Mejías-Acosta; Mayra D'Armas Regnault; Eduardo Vargas-Cano; Jesennia Cárdenas-Cobo; Cristian Vidal-Silva. Assessment of digital competencies in higher education students: development and validation of a measurement scale. *Frontiers in Education*. 2024. <https://doi.org/10.3389/feduc.2024.1497376>
12. David J. Weiss; G. Gage Kingsbury. Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*. 1984. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
13. Erick Daniel Rivera Quiñonez; Blanca Romina Barcia Rivera; Amy Linett Triviño Díaz; María Gabriela Zambrano Álvarez. Modelos matemáticos para la personalización del aprendizaje en entornos virtuales. *Multidisciplinary Journal of Sciences, Discoveries, and Society*. 2026;3(3):1-12. <https://doi.org/10.63969/18qqe109>
14. Evolución de la realidad virtual en la educación superior latinoamericana: una revisión sistemática. *Formación Universitaria*. 2025. <https://doi.org/10.4067/S0718-50062025000600035>
15. Immersive virtual reality in higher education: a systematic review of the scientific literature. *Virtual Reality*. 2025. <https://doi.org/10.1007/s10055-025-01136-x>
16. Immersive virtual reality learning environments for higher education: A student acceptance study. *Computers & Education: X Reality*. 2025. <https://doi.org/10.1016/j.cexr.2025.100105>
17. Michael Antonio Pinargote Castro; Cecibel Verónica Solórzano Ortega; María José Lozano Torres; Flavio Jesús Jiménez Mejía. Accesibilidad de la Realidad Virtual Aumentada en la Educación Universitaria: Estrategias, Desafíos y Beneficios. *Revista Científica*. 2024;9(33):252-275. <https://doi.org/10.29394/Scientific.issn.2542-2987.2024.9.33.12.252-275>
18. Learning by teaching in immersive virtual reality – Absorption tendency increases learning outcomes. *Learning and Instruction*. 2023. <https://doi.org/10.1016/j.learninstruc.2022.101716>
19. Felix J. Agbo; Solomon S. Oyelere; Jarkko Suhonen; Markku Tukiainen. Design, development, and evaluation of a virtual reality game-based application to support computational thinking. *Educational Technology Research and Development*. 2023. <https://doi.org/10.1007/s11423-022-10161-5>
20. Eduardo Campos; Irving Hidrogo; Genaro Zavala. Impact of virtual reality use on the teaching and learning of vectors. *Frontiers in Education*. 2022. <https://doi.org/10.3389/feduc.2022.965640>
21. Yue Huang; Xin Cheng; Unman Chan; Liyang Zheng; Yingwen Hu; Yueting Sun; Pengyu Lai; Jun Dai; Xuesong Yang. Virtual reality approach for orthodontic education at School of Stomatology, Jinan University. *Journal of Dental Education*. 2022;86(8):1025-1035. <https://doi.org/10.1002/jdd.12915>
22. Hanyi Jiang; Shalini Vimalasvaran; Jason K. Wang; Keng B. Lim; Shalini R. Mogali; Lorainne Tudor Car. Virtual Reality in Medical Students' Education: Scoping Review. *JMIR Medical Education*. 2022. <https://doi.org/10.2196/34860>
23. Current State and General Perceptions of the Use of Extended Reality (XR) Technology at the University of Newcastle: Interviews and Surveys From Staff and Students. *SAGE Open*. 2022. <https://doi.org/10.1177/21582440221093348>
24. Katerina Tzafilkou; Maria Perifanou; Anastasios A. Economides. Development and validation of students' digital competence scale (SDiCoS). *International Journal of Educational Technology in Higher Education*. 2022. <https://doi.org/10.1186/s41239-022-00330-0>

FUNDING

None.

CONFLICT OF INTEREST

They do not exist.

AUTHORSHIP CONTRIBUTIONS

Conceptualization: Karla Andrea Paredes Rosado

Data curation: Karla Andrea Paredes Rosado

Formal analysis: Karla Andrea Paredes Rosado

Research: Karla Andrea Paredes Rosado

Methodology: Karla Andrea Paredes Rosado

Supervision: Karla Andrea Paredes Rosado

Validation: Karla Andrea Paredes Rosado

Visualization: Karla Andrea Paredes Rosado

Writing – original draft: Karla Andrea Paredes Rosado

Writing – review and editing: Karla Andrea Paredes Rosado