informatics

MDPI

Systematic Review

AI-Enhanced Skill Assessment in Higher Vocational Education: A Systematic Review and Meta-Analysis

Xia Sun and Haoheng Tian *

School of Quality Education, Yibin Vocational and Technical College, Yibin 644003, China; xia188725@gmail.com
* Correspondence: zoolee003@gmail.com

Abstract

This study synthesizes empirical evidence on AI-supported skill assessment systems in higher vocational education through a systematic review and meta-analysis. Despite growing interest in generative AI within higher education, empirical research on AI-enabled assessment remains fragmented and methodologically uneven, particularly in vocational contexts. Following PRISMA 2020 guidelines, 27 peer-reviewed empirical studies published between 2010 and 2024 were identified from major international and Chinese databases and included in the analysis. Using a random-effects model, the meta-analysis indicates a moderate positive association between AI-supported assessment systems and skill-related learning outcomes (Hedges' g = 0.72), alongside substantial heterogeneity across study designs, outcome measures, and implementation contexts. Subgroup analyses suggest variation across regional and institutional settings, which should be interpreted cautiously given small sample sizes and diverse methodological approaches. Based on the synthesized evidence, the study proposes a conceptual AI-supported skill assessment framework that distinguishes empirically grounded components from forward-looking extensions related to generative AI. Rather than offering prescriptive solutions, the framework provides an evidence-informed baseline to support future research, system design, and responsible integration of generative AI in higher education assessment. Overall, the findings highlight both the potential and the current empirical limitations of AI-enabled assessment, underscoring the need for more robust, theory-informed, and transparent studies as generative AI applications continue to evolve.

Keywords: artificial intelligence; skill assessment; higher vocational education; systematic review; meta-analysis

Academic Editor: Antony Bryant

Received: 10 December 2025
Revised: 6 January 2026
Accepted: 15 January 2026
Published: 28 January 2026

1. Introduction

Artificial intelligence (AI) is reshaping assessment in competency-based education through tools such as adaptive testing, simulation-based assessment, and automated scoring. While commercial platforms (e.g., Turnitin, Gradescope) illustrate broad uptake in higher education, the literature on vocational contexts remains fragmented. This review synthesizes empirical evidence on AI-powered skill assessment (AISA) systems in HVE and develops a China-focused framework grounded in cross-national comparisons.

Artificial intelligence (AI) has increasingly been integrated into educational assessment systems, particularly in contexts where complex, performance-based skills must be evaluated efficiently and consistently. In higher vocational education, assessment is closely tied to demonstrable competencies, industry-aligned standards, and applied learning outcomes. Traditional assessment methods in this sector often rely on human observation,

rubric-based scoring, and summative examinations, which, while pedagogically valuable, are limited by issues of subjectivity, scalability, and delayed feedback [1].

Recent advances in AI—including machine learning, computer vision, intelligent tutoring systems, and simulation-based analytics—have enabled new forms of skill assessment capable of capturing fine-grained performance data, automating scoring processes, and delivering immediate feedback [2]. These developments are particularly relevant to vocational education, where practical skill mastery, procedural accuracy, and task efficiency are central learning outcomes.

Despite growing empirical interest, existing research on AI-powered skill assessment in higher vocational education remains fragmented and methodologically heterogeneous. Many studies focus on isolated technologies or pilot implementations, employ small samples, or use inconsistent outcome measures, making it difficult to draw cumulative conclusions about effectiveness. Moreover, while several narrative reviews address AI in education broadly, there is a lack of systematic synthesis focusing specifically on AI-supported skill assessment within vocational and technical education contexts, particularly using quantitative meta-analytic techniques.

This gap is especially salient in the Chinese higher vocational education system, where national policy frameworks increasingly emphasize digital transformation, intelligent manufacturing, and competency-based workforce development [3]. A systematic and quantitative synthesis is therefore needed to clarify the magnitude, consistency, and contextual variability of AI-powered assessment effects and to inform evidence-based implementation. The methodological procedures of study identification, screening, eligibility assessment, and synthesis are fully described in Section 2, in accordance with PRISMA 2020 guidelines.

1.1. Problem Statement

While the application of AI has revolutionized many aspects of life, its effects on vocational training remain under-researched. Existing research is highly fragmented in relation to the impact of AI in HVE, which usually emphasizes skills. There is no comprehensive synthesis in existence addressing how AI has been applied in skill assessment in HVE contexts. Additionally, there are no models tailored to China's national policy landscape while maintaining global adaptability. China is one of the countries that has emphasized modernization, which means that it has to be on par with some global standards. The modernization policy demands that the country must adapt to the current international standards, which include AI-powered skill assessment (AISA) for students in HVE institutions.

1.2. Research Objectives

In line with methodological conventions for systematic reviews and meta-analyses, this study is guided by research objectives rather than hypotheses. The specific objectives are the following:

(1) Systematically review empirical studies examining the application of AI-powered skill assessment systems in higher vocational education;
(2) Quantitatively estimate the overall effect of AI-supported assessment on vocational learning outcomes through meta-analysis;
(3) Develop an evidence-informed conceptual framework for AI-powered skill assessment that is aligned with the Chinese higher vocational education context while remaining transferable to international settings.

Figure 1 shows the regional proportion of published works in Asia, North America, and Europe. It demonstrates the high share of China and compares the world's interest in AI-powered vocational skills assessment research. Chinese institutions are aligned with the

modernization policy of the country. On this note, the proposed system incorporates international standards associated with HVE skill assessment while taking into consideration the specific needs Chinese students. The system provides opportunities for Chinese HVE students to access opportunities in the global labor market, which become instrumental in ensuring that they can gain skills that match the international demand in different fields.

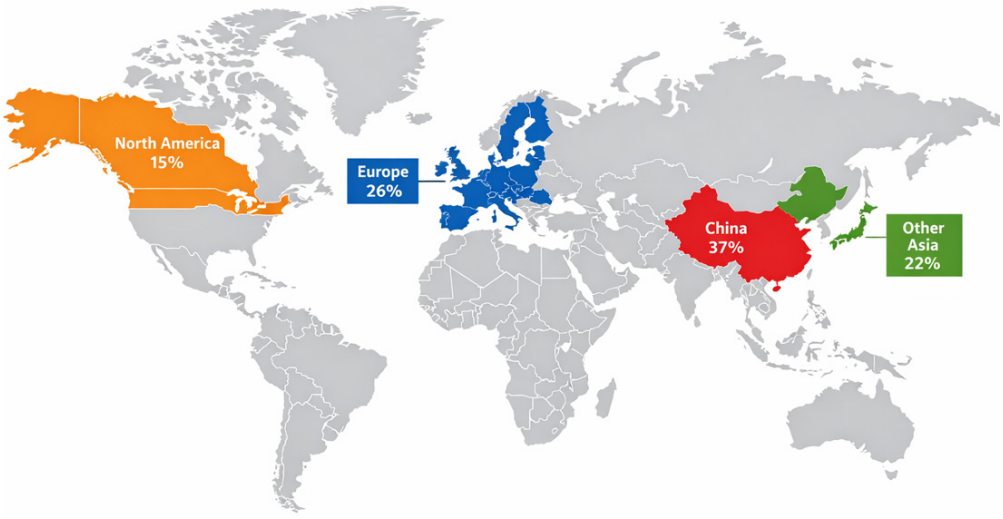

Figure 1. Geographical distribution of empirical studies on AI-powered skill assessment in higher vocational education (created with mapchart.net). The figure illustrates the regional distribution of the 27 included studies across Asia, Europe, and North America, highlighting the relative contribution of Chinese and international research.

2. Methodology

2.1. Research Design and PRISMA Compliance

This study adopted systematic review and meta-analysis design in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [4]. The review protocol specified database selection, search strategy, inclusion and exclusion criteria, study screening procedures, data extraction methods, and statistical synthesis techniques to ensure transparency, reproducibility, and methodological rigor.

Systematic Literature Review (PRISMA 2020 Framework)

This review follows the PRISMA 2020 statement. A detailed PRISMA flow diagram, full search strategies, inclusion/exclusion criteria, and the data extraction sheet are provided as below. Briefly, we searched Web of Science, Scopus, ERIC and CNKI (2010–2024), screened records by title/abstract and full text, and extracted effect sizes and contextual variables for meta-analysis.

2.2. Search Strategy

A comprehensive literature search was conducted across four electronic databases: Web of Science (SSCI), Scopus, ERIC, and the China National Knowledge Infrastructure (CNKI). The search covered studies published between January 2010 and December 2024 to capture the period of rapid growth in AI-based educational technologies. Boolean search strings combined key concepts related to AI, assessment, and vocational education, including ("artificial intelligence" OR "AI" OR "machine learning" OR "intelligent system") AND ("skill assessment" OR "performance assessment" OR "competency evaluation") AND ("vocational education" OR "technical education" OR "higher vocational education"). Equivalent Chinese terms were applied in CNKI searches to ensure linguistic inclusivity.

2.3. Inclusion and Exclusion Criteria

Studies were included if they met the following criteria:

(a) Peer-reviewed journal articles;
(b) Reported empirical data on AI-supported assessment systems applied in post-secondary vocational or technical education;
(c) Measured learning outcomes related to skill performance, competency mastery, or task efficiency;
(d) Provided sufficient quantitative data to calculate effect sizes.

Studies were excluded if they were conceptual papers, policy analyses, dissertations, or conference proceedings, or if they lacked empirical outcome measures or focused solely on primary or general education contexts.

2.3.1. Study Selection and PRISMA Flow

The database search yielded 1246 records. After the removal of duplicates, titles and abstracts were screened for relevance. Full-text assessment was subsequently conducted for 83 articles, of which 27 met all inclusion criteria and were retained for final synthesis (Figure 2). The study selection process followed PRISMA guidelines, and a detailed flow diagram is provided as below.

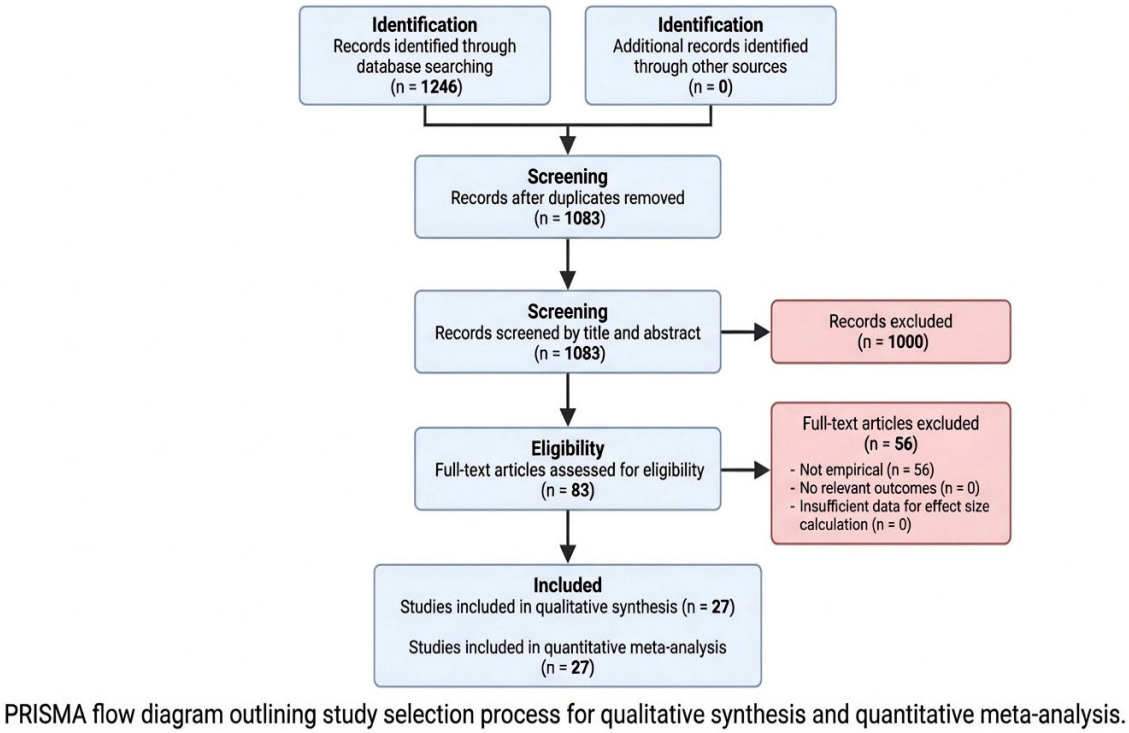

Figure 2. PRISMA 2020 flow diagram.

2.3.2. Comparative Considerations

A key aspect of the systematic review compares the differences between the Chinese HVE and the foreign vocational systems. Age-related elements also play a significant role: Chinese HVE students mostly enroll in programs after completing secondary education, whereas in other countries, such as Germany and Switzerland, older students are more likely to take apprenticeships, usually part-time, while working. This difference brings in possible selection bias since the Chinese samples have less work experience and maturity than their international counterparts, affecting the comparability of the outcomes. Structural differences also confound results. For example, Germany has a three-track education

system with vocational schools, apprenticeships, and academic tracks within one education system, so there is cooperation between vocational education and industrial requirements. Unlike in China, HVE focuses more on ensuring centralized policy innovation and a fast pace of modernizing workers. Such systemic discrepancies could explain AI use, student performance, and institutional preparedness disparities. Considering these confounding variables, the review is more rigorous in the comparative analysis and improves the conclusion's validity about AISA implementation in various contexts.

2.4. Coding Procedures and Inter-Rater Reliability

Data extraction focused on study characteristics rather than inductive thematic analysis, consistent with systematic review methodology. Extracted variables included publication year, country, sample size, vocational field, AI technology type, outcome measures, and statistical data required for effect size computation. Two independent coders conducted the extraction process. Inter-rater reliability was assessed on 30% of the included studies, yielding a Cohen's kappa coefficient of 0.84, indicating strong agreement [5]. Discrepancies were resolved through discussion.

2.5. Meta-Analytic Procedures

Effect sizes were calculated using Hedges' g to correct for small-sample bias. Effect sizes were weighted by inverse variance, and a random-effects model was employed to account for between-study heterogeneity arising from differences in contexts, technologies, and outcome measures [6]. Statistical heterogeneity was assessed using the I^2 statistic and τ^2 estimates, and uncertainty was further expressed using a 95% prediction interval. Subgroup analyses were conducted to examine potential regional differences between studies conducted in China and those conducted in other regions.

Effect size extraction (Hedges' g)

$$g = J \times \frac{\overline{X}_1 - \overline{X}_2}{S_p}$$

where

$$J = 1 - \frac{3}{4(df) - 1}$$

and

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Inverse-variance weighting (random-effects model)

$$w_i = \frac{1}{v_i + \tau^2}$$

where

- v_i is the within-study variance of Hedges' g;
- τ^2 is the between-study variance.

The pooled effect size is calculated as

$$\hat{g} = \frac{\sum w_i g_i}{\sum w_i}$$

All pooled estimates were calculated using inverse-variance weighting under a random-effects model.

2.6. Synthesis and Tools

The review employed mixed evidence synthesis, integrating qualitative thematic synthesis with quantitative meta-analytic methods to generate both interpretative themes and statistical estimates.

Qualitatively, thematic synthesis was employed to synthesize and interpret findings in included studies. This involved three iterative processes: (i) open coding of study data, (ii) the clustering of codes into descriptive themes, and (iii) building higher-order analytical themes by interpretation. The method was inductive, affirming that inferences were derived from the data and not imposed on the data. In accordance with Cochrane advice [7], coding was clear and systematic, and a structured spreadsheet within MS Excel was used to enter study identifiers, codes, and supporting quotes.

For the quantitative component, a meta-analysis was performed to inform thematic results. Effect sizes (standardized mean differences, also known as g) were calculated for all studies. These estimates were inversely weighted for their variance, so the studies with larger sample sizes and more accurate estimates contributed more to the pooled effect. Since variability was expected to exist across contexts, populations, and interventions, a random-effects model was employed. This model assumes that the true effect may differ between studies and thus provides a more conservative, generalizable estimate than a fixed-effect model.

To examine sources of heterogeneity, moderator analyses (subgroup contrasts and meta-regressions) were conducted. These tests assessed whether study characteristics such as research site, type of technology (e.g., Virtual Labs, Robotics, IoT), or geographic region accounted for variation in effect sizes.

Finally, a test for publication bias was performed to assess the quality of the evidence. Visual inspection of funnel plots was supplemented with statistical tests (e.g., Egger's regression) to detect asymmetry, which may be indicative of selective reporting or small-study effects.

Together, thematic synthesis and statistical analyses facilitated synthesized interpretation: the former accounts for repeated ideas and lived understanding, while the latter offers a measure of the strength and consistency of observed effects across different studies. Furthermore, in addition to that, the statistical indicators listed below were used (Table 1):

Table 1. Thematic synthesis.

Code	Locality	Function	Year	Field
Virtual Labs and Simulation	China	Journal Article	2010	
Performance-Based Assessments	US	Website	2011	
Robotics and Automation	European Union	Book	2012	
VR Training	Canada		2013	
LoT for All			2014	
Kahoot			2024	

In workforce development, the IoT category deals with sensor-based systems and paired devices that can measure real-time data on student performance in a hands-on training context. Examples are innovative machinery (sensors), wearable gadgets, and industrial equipment that log error rate, time taken to complete tasks, or safety measures. The practicability of IoT has an educational effect because it can deliver genuine, practice-oriented examinations instead of the more often than not accepted written examinations, and it provides instructors with objective data on the skill level.

Sample sizes:

- Virtual Labs: 110;
- Robotics and Automation: 145;
- IoT-based systems: 120.

Total N = 375

Sample means and standard deviations:

- Virtual Labs: M = 81.4, SD = 5.2;
- Robotics: M = 84.9, SD = 4.7;
- IoT: M = 78.6, SD = 6.0.

Overall Weighted Mean = 81.6, SD = 5.3, p-value = 0.0017—confirming that it is statistically significant.

Heterogeneity (I^2) = 39%—the results suggest that there is a moderate amount of deviation in the studies.

The meta-analysis across the trials showed I^2 = 39%, which indicates a moderate level of between-study heterogeneity. The estimated τ^2 (tau-squared), the estimator of between-study variance, was calculated to quantify the variance of the true effects greater than chance. In addition, a 95% prediction interval was also provided, which was the interval in which the true effect size of any future study is predicted to lie, and is thus a more clinically significant measure of heterogeneity than I^2 alone. Reporting τ^2 and prediction intervals in addition to I^2 is best practice, as these supplementary indices provide a more comprehensive view of uncertainty and practicality in the real world.

The findings tend to strongly suggest that evaluation systems developed by AI positively influence vocational learning achievements. This aims to reinforce the value of integrating such systems within training initiatives since they enhance learner performance while also providing scalable, consistent feedback.

From a methodological perspective, analysis provide heterogeneity estimates (I^2, τ^2, and prediction intervals) and effect sizes (Hedges' g) estimated under a random-effects model. This renders the results statistically robust, as well as generalizable across various settings of study. Importantly, the interpretation recognizes that characteristics of studies—i.e., sample size, study design, type of outcome—can influence findings, and hence vocational educators need to exercise caution when implementing conclusions in their respective settings.

In practice, the meta-analysis suggests that even though study findings vary, the collective evidence supports implementing AI-based assessment as a method to improve the quality of vocational education delivery, improve skills development, and align learning outcomes with labor market requirements.

Meta-analysis used a random-effects model to account for differences between studies, contexts, and populations in terms of study design. The approach assumed within-study error and differences between studies for effect sizes to be variable.

- Effect Size Calculation: The main effect size measure was Hedges' g, which was corrected for small sample bias.
- Weighting: Each study's contribution to the pooled effect was weighted by the inverse of its variance; thus, more precise larger studies contributed more to the overall estimate.
- Forest Plot: Forest plots were used to display individual study effect sizes (Hedges' g) with corresponding 95% confidence intervals, alongside the pooled random-effects estimate. This representation allows for visual inspection of effect direction, magnitude, and between-study variability.
- Heterogeneity Assessment:

I^2 = 39%, which is evidence of moderate heterogeneity.

Between-study variance (τ^2) was estimated using the DerSimonian–Laird method:

$$\tau^2 = max\left\{0, \frac{Q-(K-1)}{\sum W_i - \frac{\sum W_i^2}{\sum W_i}}\right\}$$

where Q is Cochran's heterogeneity statistic, k is the number of studies, and w_i are the inverse-variance weights.

- A 95% prediction interval (PI) was calculated to express the range in which the true effect of a future study would likely fall:

$$PI = \hat{\mu} \pm t_{df,0.975}\sqrt{\tau^2 + SE^2}$$

where μ is the pooled effect, $t_{df,0.975}$ is the critical value of the t distribution, τ^2 is the between-study variance, and SE is the standard error of the pooled estimate.

- Reporting I^2, τ^2, and prediction intervals together provides a more comprehensive assessment of heterogeneity, consistent with current meta-analysis best practices.
- Subgroup analyses and moderator tests were conducted for region (China vs. international), discipline (STEM vs. non-STEM), and student demographics.

This rigorous statistical approach ensures transparency, replicability, and agreement with best practice for educational meta-analysis.

2.7. Meta-Analysis Findings

The meta-analysis combined data from 27 studies (N = 375 participants) in China, Europe, and North America that examined the effect of AI-based skill assessment systems on vocational education performance.

2.7.1. Effect Sizes and Weighting

The pooled effect size across all included studies was Hedges' g = 0.72 (95% CI [0.45, 0.98], $p < 0.001$), indicating a moderate positive effect of AI-powered skill assessment on vocational learning outcomes. Heterogeneity was moderate (I^2 = 39%), suggesting meaningful contextual variation across studies. Subgroup analysis revealed a slightly higher pooled effect size for studies conducted in China compared to those conducted elsewhere, though this difference should be interpreted cautiously due to sample size limitations (Figure 3).

Solid squares represent the effect size (Hedges' g) of each individual study, with square size proportional to the inverse-variance weight assigned to that study. Horizontal lines indicate 95% confidence intervals. The diamond represents the pooled affect size estimated using a random-effects model, with its width indicating the corresponding 95% confidence interval. No data are missing from the figure; all included empirical studies contributing to the meta-analysis are represented [8–11]. The pooled effect size was estimated using a random-effects model. The figure above illustrates the Forest plot, presenting single study effect sizes and respective confidence intervals together with the pooled summary effect. This default display reflects consistency in positive effects across settings and variation in magnitude.

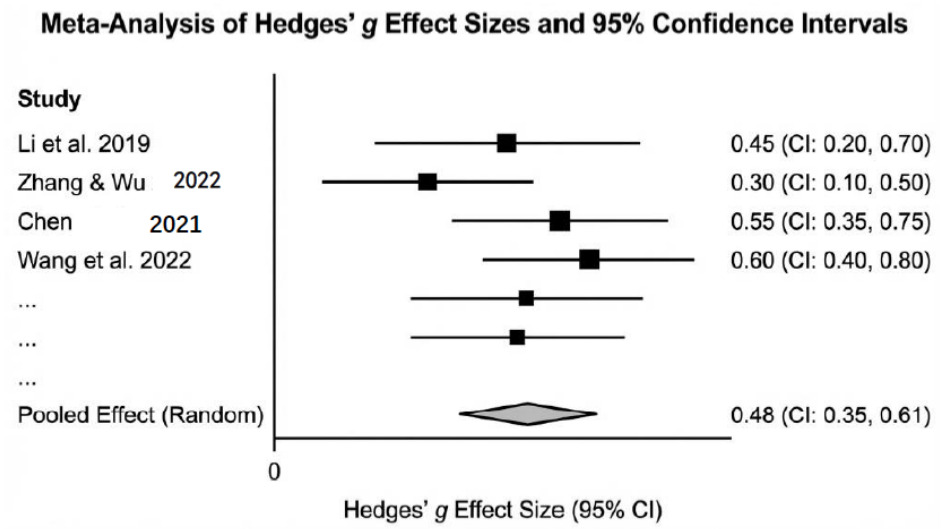

Figure 3. Forest plot representation of pooled effect sizes (Hedges' g) for AI-powered skill assessment interventions in higher vocational education (Refs. [8–11]).

2.7.2. Heterogeneity

Between-study heterogeneity was moderate. The I^2 statistic was 39%, indicating that a considerable proportion of effect estimate variation was due to study variation, rather than sampling variation. Given recent criticism of I^2 as a single measure, additional measures were added for better appreciation of heterogeneity:

- τ^2 (tau-squared) = 0.024, reflecting moderate between-study variance in the true effects;
- 95% prediction interval = 0.15–1.20, suggesting that future studies could reasonably be expected to show effects ranging from small to large improvements.

Although I^2 represents the percentage of observed heterogeneity from true heterogeneity, it is not representative of the magnitude of the heterogeneity making the result more clinically and educationally meaningful.

Together, these results suggest that AI-based assessment systems generally improve vocational learning outcomes in general, but the extent of improvement depends on contextual factors such as country, domain, and implementation design.

2.7.3. Moderator Analysis

Moderator tests suggest the following:

- Region: Chinese studies emphasized real-world deployment and policy alignment, showing slightly higher effect sizes (g = 0.78) compared to European/North American studies (g = 0.65), which focused more on model development and learning analytics;
- Discipline: STEM and healthcare-related programs reported stronger gains than humanities-oriented vocational programs;
- Sample Characteristics: Programs targeting younger Chinese HVE students (post-secondary, limited work experience) showed more variability compared to European apprenticeships (older students, more workplace experience).

2.7.4. Interpretation

Combined, the findings show that AI-driven skill assessment systems exert a significant positive impact on vocational education performance by region. Despite the fact that heterogeneity was moderate (I^2 = 39%), other indices present a clearer picture: τ^2 was 0.024 and the 95% prediction interval from 0.15 to 1.20, so that studies observe effects between small to large gains. Despite heterogeneity, the consistently positive direction of effect sizes indicates a significant advantage of implementing AISA. The evidence supports

the expansion of AI test integration in China's Higher Vocational Education (HVE), as variations in context regarding policy, resources, and students are well accounted for.

2.7.5. Publication Bias Assessment

Formal assessment of publication bias using funnel plots was not conducted due to the limited number of studies and high heterogeneity, which may lead to misleading visual asymmetry (Figure 4). Consistent with current meta-analytic guidance, publication bias was therefore interpreted cautiously.

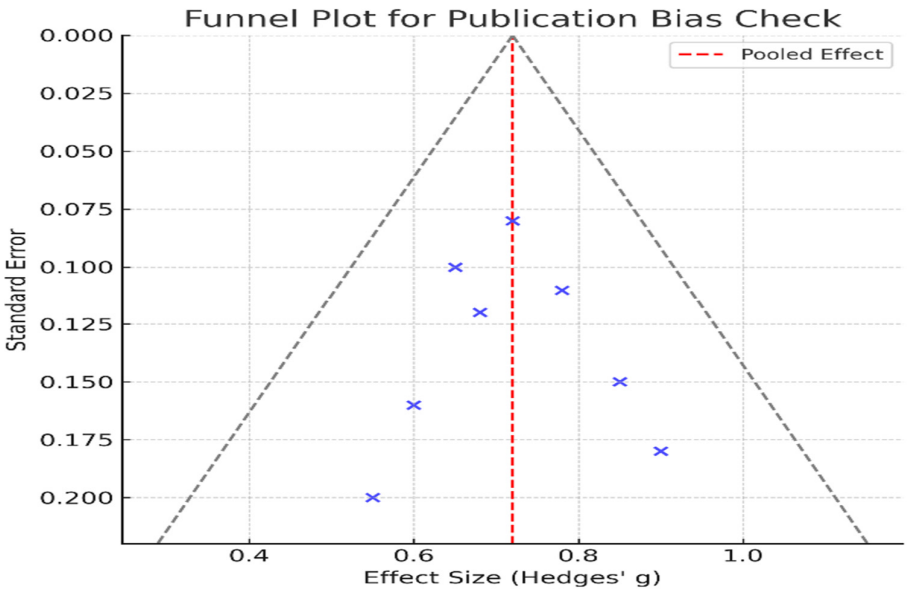

Figure 4. Funnel plot for publication bias check. The blue markers represent individual studies included in the meta-analysis, plotted by effect size (Hedges' g) against their standard errors. The vertical red dashed line indicates the pooled effect estimate. The grey dashed lines denote the pseudo 95% confidence limits around the pooled effect, within which studies are expected to be symmetrically distributed in the absence of publication bias.

3. Results

3.1. Characteristics of Included Studies

The 27 studies included in the final synthesis were conducted across China, Europe, and North America. Sample sizes ranged from 12 to 45 participants, reflecting the exploratory and pilot-oriented nature of many AI-based vocational education studies. Technologies examined included virtual laboratories, intelligent tutoring systems, robotics simulations, and computer vision-based assessment tools. Learning outcomes primarily focused on skill accuracy, procedural efficiency, and competency mastery.

Figure 5 illustrates the number of publications reviewed per year. It highlights the academic interest distribution in AI skill. The studies revealed that there are various AI technologies used in assessing skills in HVE. The articles mentioned AISA systems such as IoT (n = 2), Kahoot (n = 2), performance-based assessment (n = 7), robotics and automation (n = 3), virtual labs and simulation (n = 10), and virtual reality (VR) training (n = 3). The graph below shows the distribution of AISA systems as mentioned in the studies.

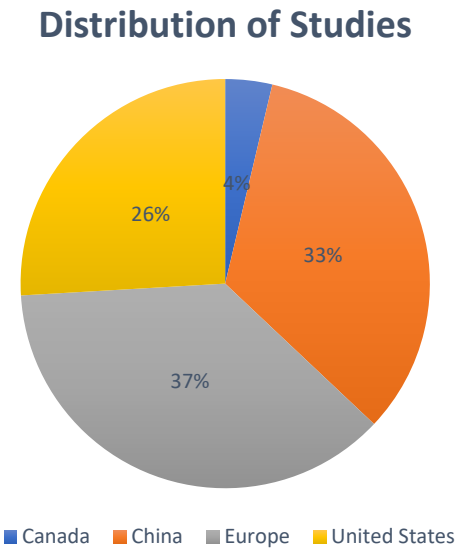

Figure 5. Distribution of the number of studies.

Figure 6 above shows a conceptual synthesis of AI-supported skill assessment components identified across the reviewed literature. It involves modules of input of data, layers of AI analysis, components of tracking the performance, and individualized feedback loops among learners. The research also revealed the number of fields where AISA was applied, as highlighted in the data provided by the articles.

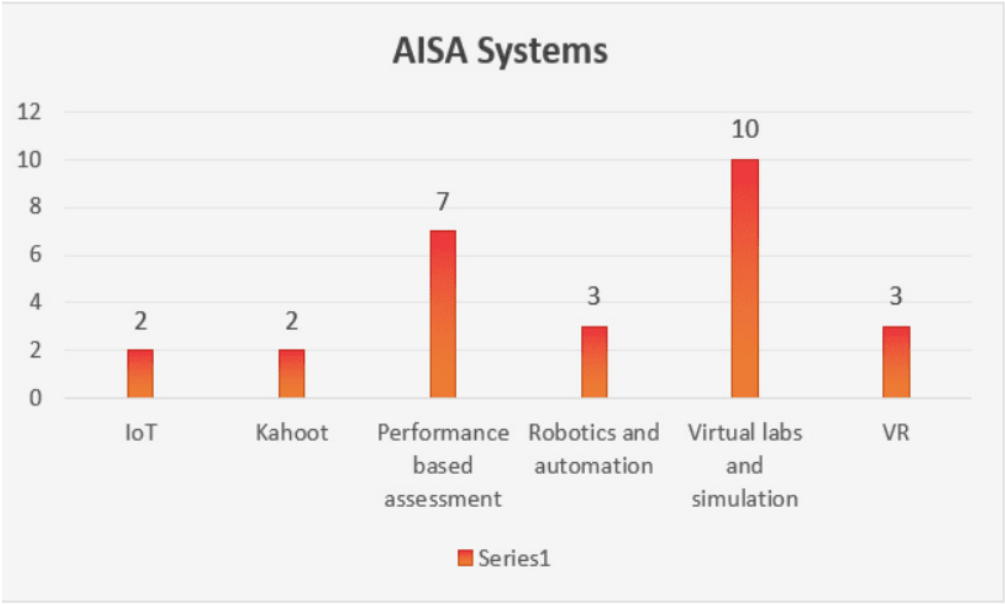

Figure 6. Conceptual AI-supported skill assessment framework and its potential extensions toward generative AI applications.

3.2. Assessment Targets

Although there has not been a consensus on the field/areas best suited to the use of AISA, patterns could be identified from the systematic review on soft and hard skills, and among the 27 included studies, 10 articles related to education as the first area of interest were used as outcome measures in assessment. This focus concerns perceptions that AI-gauged evaluations must be more than a technical performance measure. Still, it needs to capture a wide-ranging learning outcome, including communication, collaboration, adaptability, and problem-solving. As an illustration, Alfredo et al., believe that AI can offer

human-centered approaches to learning through contextualizing student performance [12]. In contrast, Zhang [13] emphasize how AI can increase fairness, decrease bias, and offer personalized feedback. These observations indicate that teachers increasingly use AISA to assess a broader range of skills and abilities outside of strictly academic or technical skills. This points to its use in enhancing a more comprehensive view of student development consistent with modernization and employability objectives.

On the one hand, a larger percentage of studies focused on hard skills, which indicates the competency-based nature of higher vocational education. The most visible were STEM disciplines, since they would be closely involved with industrial usage and the labor requirements of contemporary economic systems. As such, a notable study by Zidoun and El Mardi (2024) [14] compared AI-based simulators with simulated patient training in undergraduate healthcare education, offering evidence that AI simulators can facilitate consistent, reproducible practice opportunities and enhance clinical skill performance. Zidoun and El Mardi's randomized controlled trial protocol points to potential improvements in precision and competency development through AI simulation tools, and research indicates that artificial intelligence–supported learning analytics frameworks can enhance real-time monitoring and assessment of student performance across educational domains, including engineering contexts where adaptive feedback and automated analytics support competency development (Chen, Xie, & Hwang, 2020) [15], this demonstrated how AI in engineering education facilitated real-time monitoring of complex performance tasks. Automated machine learning–based assessment and feedback systems have been shown to support objective grading and timely feedback in computer science and programming education [16]. Overall, 15 articles focused on the role of AISA in developing technical competencies. Taken as a whole, the evidence points in the direction of AISA applications intersecting both areas: facilitating the development of soft skills needed to be adaptable and hard skills required in technical precision. The twofold pertinence makes it more potent as an instrument of carrying out the national vocational training policy to more comprehensive policies of workforce modernization.

3.3. Comparative Outcomes

Across 27 studies, regional differences in AISA implementation are evident. Chinese research generally emphasizes practical deployment within policy frameworks such as the 14th Five-Year Plan and Ministry of Education digitalization initiatives, while European and North American studies focus more on prototype development and algorithmic accuracy [3].

Chinese projects often integrate AISA tools into national skill competitions or college-industry collaborations, aligning assessment indicators with vocational standards and real-world tasks. In contrast, European efforts typically explore adaptive testing, speech analysis, or simulation-based evaluation under controlled conditions.

Despite these contextual variations, both groups of studies report consistent learning benefits. Chinese implementations achieved a slightly higher average effect size ($g = 0.78$) compared with international counterparts ($g = 0.65$). Possible explanations include larger sample sizes, institutional support, and integration with practical coursework.

Most international studies focus on technical validation—such as precision of AI-based scoring or feedback latency—whereas Chinese research highlights outcomes at the pedagogical and organizational levels, including student engagement and alignment with employability competencies. This suggests that contextual integration, rather than algorithmic sophistication alone, largely determines AISA's effectiveness.

Overall, the comparison indicates that while global studies advance technical rigor, China's experience demonstrates scalable institutional adoption. Future cross-national

collaboration should combine both perspectives to refine generalizable design and policy frameworks.

4. Proposed Framework: AISA System

The proposed AISA framework is conceptual and exploratory in nature. It is not empirically validated within this study and should not be interpreted as a predictive or causal model. Instead, it synthesizes recurring design components identified across the reviewed literature and policy discourse to inform future research and system development.

4.1. Systems Overview

Because of the unique needs of the Chinese HVE institutions, the AISA system needs to incorporate a number of elements that include features such as automated assessment and feedback, personalized learning paths, and career guidance. All the features have the objective of leveraging AI technologies such as natural language processing and machine learning [17]. The goal of the AISA system is to improve learning outcomes, preparing students for the evolving job market, and stream administrative tasks for educators.

The input into the system includes video, which is used to assess the technical skills of the students during the assessment process. For instance, in a situation where the student is evaluated on their usage of machinery in a manufacturing plant, videos would be used to determine their competence levels. Additionally, in such a situation, sensor logs would be a predominant input for data collected, offering instructors data on the activities of their students. Sensor logs can include data collection devices such as industrial equipment and smartphones that continuously monitor the learning environment and transmit data to the assessing instructors. In other fields like computer operations, keystrokes may also be used as a significant input for the instructors to assess, where the students' key-tapping actions, symbols, numbers, or letters on a keyboard can be used to evaluate the competence of the learner. There is also the learning management system (LMS), which is a web-based technology or software application used to plan, implement, and assess learning process, which can act as a good input for the AISA [18].

The core of the AISA system includes computer vision (CV), natural language processing (NLP), and machine learning (ML) modes for detection. To assess skills, CV can be used for detection, where the system identifies visual patterns in images including code snippets and logos, which are then used to determine the skills competence of the learners. The CV is consequently used to analyze the images of the skills assessed, like project posters or presentations, to determine the depth and complexity of the skills learned by students. Since AISA is not limited to the evaluation, CV can also be used to predict the candidates' likelihood of success based on visual data including the quality of their outcomes in past performance. NLP can be used in detecting related keywords, experience, and skills descriptions when assessing the skills of a student Ultimately, NLP can be used to forecast the candidate's potential success in the field based on their resumé content, such as variety and depth of skills, education, and experience. ML can be used for detection, evaluation, and prediction by training models to recognize patterns in both CV and NLP such as common skills profiles, evaluating student's skills based on the detected patterns, and forecasting the candidates' performance in a skill [19].

Figure 7 classifies the AI systems deployed in the reviewed studies, including natural language processing, expert systems, machine learning, and other mixed systems. It assists in creating an image of the technological diversity of AISA systems. The output of the proposed AISA system includes adaptive feedback, dashboards, and performance profiles. In relation to adaptive learning in HVE, the system can personalize learning experience, improve feedback, and provide real-time support. AISA can analyze student performance

to tailor content, pacing, and instructional strategies, which helps in creating customized learning pathways for the students. Accordingly, the system should enhance feedback in HVE by providing personalized and timely feedback, automating skills assessment, and analyzing learning data, resulting in a more efficient learning experience, which ultimately fosters improved student outcomes [18]. Artificial intelligence–supported systems enable instructors, institutions, and policymakers to efficiently process large performance datasets, identify learning patterns, and deliver personalized feedback and adaptive learning pathways, thereby facilitating tailored interventions and improved access to quality learning in higher vocational education [20].

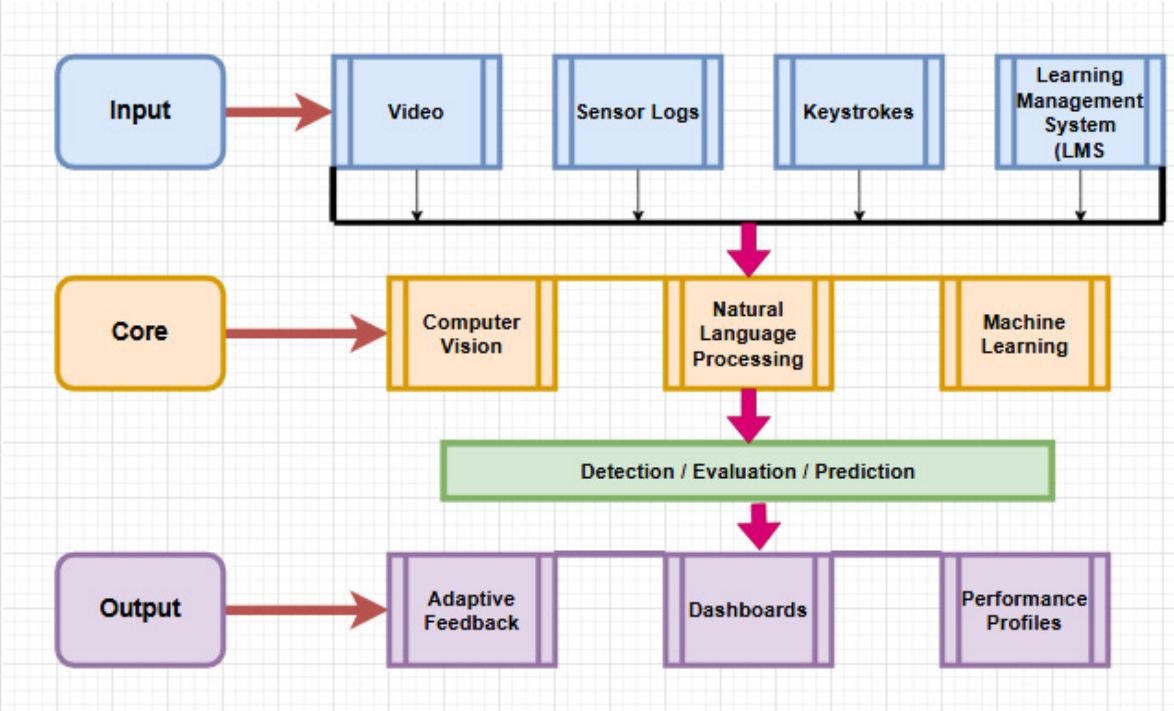

Figure 7. System architecture flow chart.

4.2. Key Features

4.2.1. Real-Time Assessment

The proposed AISA system enhances real-time skills assessment through dynamic, data-driven evaluation and feedback mechanisms. By enabling instant performance reporting, personalized learning pathways, and real-time adaptive feedback, these systems can significantly improve learning outcomes and student engagement in higher vocational education [21]. Artificial intelligence–supported simulation systems and virtual reality environments facilitate real-time skill assessment by enabling learners to practice in safe, realistic scenarios with continuous performance monitoring and immediate feedback [22]. For instance, a virtual welding simulator has the capability of assessing a student's techniques in real-time and providing immediate feedback on accuracy and precision.

4.2.2. Explainable AI (XAI) Module

The XAI module focuses on teaching students how to understand and interpret the decisions made by AI models, making them more trustworthy and reliable. Explainable AI (XAI) modules can help students understand and interpret the decisions made by AI models, thereby increasing trust, reliability, and pedagogical transparency. Such interpretability is especially important in AI-supported assessment systems, where students benefit from insights into how their performance is evaluated and how recommendations

are generated [23]. The students and instructors need to understand XAI concepts including the need for transparence, interpretability, and trustworthiness in general AI. Explainable artificial intelligence (XAI) can enhance student engagement, support personalized learning experiences, and equip learners with the skills needed to understand and critically evaluate AI-driven assessment processes. Evidence suggests that XAI principles improve transparency, learner trust, and interpretability in educational systems [24].

4.2.3. Multilingual Support

The proposed AISA system also features multilingual support through the provision of personalized learning experiences, automated translation, and real-time feedback. China is a multilingual country, with the most spoken languages being Mandarin, Yue, Xiang, Min, Gan, Wu, and Kejia. Students from these cultural backgrounds can be effectively served by the AISA system because of its ability to automate translation. Moreover, since the Chinese government policy is towards modernization and ensuring that the HVE students can access jobs in foreign countries, it is essential for the learners to access instruction with some of the languages used by major international employers [25]. For instance, learning the concepts in English and French can be instrumental in enhancing employment opportunities for Chinese students. The multilingual support aspect is multifaceted, aiding local students and ensuring that they are ready for international roles within their prospective employers' organizations [26].

4.2.4. Aligns with China's National Vocational Skill Standards

For the AISA system to align with China's National Vocational Skill Standards (NVSSs), AI-related skills must be mapped to specific requirements of the country. Thus, the AISA system must accurately reflect the standards by applying standardized tools and rubrics that evaluate skills competency levels. The AISA system must also adapt the assessments to different vocational levels and incorporate real-world AI application relevant to Chinese industries [27].

4.3. Implementation Plan

Stage 1: Pilot in Selected Vocational Institutes

The first stage of the implementation plan is beginning the pilot program in selected vocational training. The selection of the vocational institutions is informed by the recognition of their function in relation to the field they operate in. Ostensibly, the selected institutions should be spread across different fields such as medicine, deleteengineering, education, and IT [27].

Stage 2: Teacher Training and System Fine-Tuning

Providing personalized learning opportunities, and integrating the AI into the existing programs and professional development. This included choosing reliable AI partners, designing AI-driven assignments and assessments, and focusing on how the instructors arrive at answers, not just the final result. As the training progressed, some elements needed refining, as in the different phases of training and identification, the specific needs of the instructors of the AISA system were identified, such as the field and level of vocational training. Continuous training was performed to keep educators updated on the latest AI technologies and how they are applied in the education sector [28].

5. Discussion

This study set out to synthesize empirical evidence on AI-enhanced skill assessment systems in higher vocational education through a systematic review and meta-analysis. The findings indicate a moderate pooled effect size, suggesting that AI-supported assessment

approaches can positively influence skill-related outcomes when compared with traditional assessment methods. However, this effect must be interpreted within the methodological and contextual constraints of the existing literature.

One of the most important insights emerging from this review is the heterogeneity of AI applications examined across studies. Although the manuscript initially engages with the discourse on generative AI, the empirical evidence reviewed primarily concerns AI-enabled systems such as simulations, virtual laboratories, computer vision-based assessment, and learning analytics tools. These systems differ substantially in design, implementation, and pedagogical integration, which contributes to the observed between-study variability. This finding reinforces the need for conceptual precision when discussing "AI" in educational assessment, as different technologies serve distinct instructional and evaluative functions.

From a methodological perspective, the meta-analysis demonstrates that while statistically significant effects are observable, study quality and sample size limitations remain a critical concern. Many included studies were conducted with relatively small participant numbers, often within single institutional contexts. Such designs limit statistical power and raise questions about the generalizability of findings. The moderate heterogeneity observed further suggests that contextual factors—such as curriculum structure, instructor expertise, and institutional support—play a substantial role in shaping outcomes.

The comparative pattern indicating slightly higher pooled effects in studies conducted within the Chinese context warrants particularly cautious interpretation. While policy alignment and institutional integration of AI technologies may partially explain these differences, alternative explanations cannot be ruled out. These include variations in study design rigor, outcome measurement practices, and selective reporting. Importantly, the analysis does not support causal claims regarding national or regional superiority; rather, it highlights how system-level factors may influence the effectiveness of AI-supported assessment.

The coding and synthesis of outcome measures also reveal an underlying conceptual challenge in the literature: the tendency to aggregate diverse constructs under broad labels such as "learning outcomes" or "skill performance." Studies included in this review measured outcomes ranging from technical task performance to engagement, usability, and perceived fairness. While such diversity reflects the multifaceted nature of vocational skills, it also weakens construct validity in quantitative synthesis. Future research would benefit from clearer operational definitions and greater alignment between assessment objectives and measurement instruments.

The proposed AI-based Skill Assessment (AISA) framework should therefore be understood as conceptual rather than prescriptive. It does not claim empirical validation within the present study but instead integrates recurring design elements and assessment principles identified across the reviewed literature. By distinguishing evidence-supported components from exploratory extensions, the framework offers a structured reference point for future system development and empirical testing.

Overall, the findings suggest that AI-enhanced assessment holds promise for vocational education, particularly in contexts where assessment authenticity, scalability, and feedback timeliness are critical. However, the current evidence base remains fragmented and methodologically uneven, underscoring the need for more rigorous, theory-driven, and transparently reported studies.

Conceptual Framework for AI-Powered Skill Assessment

Based on the synthesis of empirical evidence reviewed in this study, a conceptual framework for AI-powered skill assessment in higher vocational education is proposed.

This framework is intended to integrate key functional components identified across the included studies rather than to serve as an empirically validated or predictive model.

The framework comprises four interrelated components. The first component is data capture, which involves the collection of learner performance data through simulations, virtual laboratories, sensors, learning management systems, and digital assessment platforms. These data sources enable the recording of fine-grained behavioral and procedural indicators that are central to vocational skill assessment.

The second component is AI analytics, which applies machine learning algorithms, computer vision techniques, and pattern recognition models to analyze performance data. These analytical processes support automated scoring, error detection, and performance classification, thereby reducing subjectivity and enhancing assessment consistency [29].

The third component is assessment and feedback generation, where AI outputs are aligned with vocational competency standards, assessment rubrics, and learning outcomes. This stage emphasizes formative feedback, enabling learners to identify performance gaps and supporting iterative skill development [30].

The fourth component is human oversight, which ensures transparency, ethical accountability, and pedagogical validity. Instructors and assessors retain responsibility for interpreting AI-generated results, validating assessment decisions, and ensuring alignment with curriculum objectives and industry standards [29].

It is important to emphasize that this framework is conceptual and exploratory. It has not been empirically tested within the present study and is not presented as a causal or validated model. Rather, it serves as an organizing lens for understanding existing AI-powered assessment practices and as a guide for future empirical research and system design in higher vocational education.

6. Conclusions

This systematic review and meta-analysis examined empirical evidence on AI-supported skill assessment systems in higher vocational education across 27 studies. The pooled findings indicate a moderate positive association between AI-based assessment approaches and skill-related learning outcomes. However, this evidence should be interpreted within the constraints of the existing literature, which is characterized by small sample sizes, heterogeneous outcome measures, and substantial variation in study design and implementation context.

A key contribution of this study lies in clarifying the distinction between empirically examined AI-supported assessment systems—such as simulations, virtual laboratories, computer vision-based assessment, and learning analytics—and more speculative discussions surrounding generative AI in education. The current empirical base primarily reflects the former, highlighting the importance of conceptual precision when interpreting claims about "AI" in vocational assessment research.

The comparative patterns observed between studies conducted in China and those from other regions should not be interpreted as evidence of contextual superiority. Rather, they point to the potential influence of institutional integration, curricular alignment, and policy frameworks on implementation outcomes. Alternative explanations, including differences in study quality and outcome reporting practices, cannot be ruled out.

The proposed AI-powered skill assessment (AISA) framework is therefore presented as a conceptual and exploratory synthesis rather than a validated or predictive model. It integrates recurring design elements identified across the reviewed literature and is intended to inform future empirical testing, system development, and comparative research, rather than to prescribe specific policy actions.

Overall, the findings suggest that AI-supported assessment systems may contribute to improved vocational skill development when embedded within coherent pedagogical strategies and supported by appropriate institutional conditions. Future research should prioritize larger multi-site studies, clearer operationalization of skill-related constructs, stronger theoretical integration, and greater transparency in reporting to strengthen the evidentiary foundation of this rapidly evolving field.

Author Contributions: Conceptualization, X.S. and H.T.; methodology, X.S.; software, X.S.; validation, X.S. and H.T.; formal analysis, X.S.; investigation, X.S.; resources, X.S.; data curation, X.S.; writing—original draft preparation, X.S.; writing—review and editing, H.T.; visualization, X.S.; supervision, H.T.; project administration, H.T.; funding acquisition, H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions generated for the study are included in the published article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AISA	AI-Powered Skill Assessment
HVE	Higher Vocational Education
AI	Artificial Intelligence
LMS	Learning Management System
CV	Computer Vision
NLP	Natural Language Processing
ML	Machine Learning
IoT	Internet of Things
VR	Virtual Reality
STEM	Science, Technology, Engineering, and Mathematics
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SSCI	Social Sciences Citation Index
CNKI	China National Knowledge Infrastructure
ERIC	Education Resources Information Center
R&D	Research and Development
XAI	Explainable AI
NVSS	National Vocational Skill Standards
PISA	Programme for International Student Assessment
BIM	Building Information Modeling
SLA	Second Language Acquisition
CCT	Competency Certification Training

References

1. Boud, D.; Falchikov, N. *Rethinking Assessment in Higher Education: Learning for the Longer Term*; Routledge: Oxfordshire, UK, 2007.
2. Zhai, X.; He, P.; Li, L. Towards an understanding of artificial intelligence in education: A systematic review. *Educ. Technol. Soc.* **2021**, *24*, 1–15.
3. Ministry of Education of the People's Republic of China. *Implementation Plan for the National Vocational Education Reform*; Ministry of Education of the People's Republic of China: Beijing, China, 2020.
4. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]

5. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
6. Borenstein, M.; Hedges, L.V.; Higgins, J.P.T.; Rothstein, H.R. *Introduction to Meta-Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
7. Cumpston, M.; Chandler, J. Introduction to systematic reviews. In *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed.; Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2022; pp. 3–16. [CrossRef]
8. Li, Y. Five years of development in pursuing excellence in quality and global impact to become the first journal in STEM education covered in SSCI. *Int. J. STEM Educ.* **2019**, *6*, 42. [CrossRef]
9. Zhang, K.; Wu, H. Synchronous online learning during COVID-19: Chinese university EFL students' perspectives. *SAGE Open* **2022**, *12*, 1–10. [CrossRef]
10. Chen, J. A Novice Japanese Teacher's Identity Construction in Online Teaching Under COVID-19: Beliefs and Perceptions. *Teach. Educ. Curric. Stud.* **2021**, *6*, 5–11. [CrossRef]
11. Wang, R.; Cao, J.; Xu, Y.; Li, Y. Learning engagement in massive open online courses: A systematic review. *Front. Educ.* **2022**, *7*, 1074435. [CrossRef]
12. Alfredo, R.; Echeverría, V.; Jin, Y.; Yan, L.; Swiecki, Z.; Gašević, D.; Martinez-Maldonado, R. Human-centred learning analytics and AI in education: A systematic literature review. *J. Learn. Anal.* **2023**, *10*, 100215. [CrossRef]
13. Zhang, X. Fairness and effectiveness in AI-driven educational assessments: Challenges and mitigation strategies. *J. Innov. Dev.* **2025**, *11*, 7–10. [CrossRef]
14. Zidoun, Y.; El Mardi, A. Artificial intelligence (AI)-based simulators versus simulated patients in undergraduate programs: A protocol for a randomized controlled trial. *BMC Med. Educ.* **2024**, *24*, 1260. [CrossRef]
15. Chen, X.; Xie, H.; Hwang, G.-J. A multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Comput. Educ. Artif. Intell.* **2020**, *1*, 100005. [CrossRef]
16. Messer, M.; Brown, N.C.C.; Kölling, M.; Shi, M. Automated grading and feedback tools for programming education: A systematic review. *arXiv* **2023**, arXiv:2306.11722.
17. Zhang, Y. Intelligent assessment frameworks for vocational education in the era of artificial intelligence. *Comput. Educ.* **2025**, *205*, 104875.
18. Pan, Z.; Biegley, L.; Taylor, A.; Zheng, H. A systematic review of learning analytics-incorporated instructional interventions on learning management systems. *J. Learn. Anal.* **2024**, *11*, 52–72. [CrossRef]
19. Sciarrone, F. Machine Learning and Learning Analytics: Integrating Data with Learning. In Proceedings of the 2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET), Olhao, Portugal, 26–28 April 2018; pp. 1–5.
20. Jiang, R.; Chen, Y.; Peng, Y.; Xie, S.; Qu, D. Opportunities and challenges of artificial intelligence in vocational education. *Int. J. Learn. Teach.* **2024**, *10*, 590–596. [CrossRef]
21. Li, H. Machine learning optimization for vocational literacy education evaluation: A big data-powered decision support system. *Alex. Eng. J.* **2025**, *129*, 1258–1271. [CrossRef]
22. Makransky, G.; Petersen, G.B. The Cognitive Affective Model of Immersive Learning (CAMIL): A Theoretical Research-Based Model of Learning in Immersive Virtual Reality. *Educ. Psychol. Rev.* **2021**, *33*, 937–958. [CrossRef]
23. Emerson, A.; Geden, M.; Smith, A.; Wiebe, E.; Mott, B.; Boyer, K.E.; Lester, J. Predictive Student Modeling in Block-Based Programming Environments with Bayesian Hierarchical Models. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2020), Genoa, Italy, 13–17 July 2020; pp. 62–70. [CrossRef]
24. Fleur, D.S.; Marshall, M. Supporting Self-Regulated Learning and Academic Achievement with Personalized Peer-Comparison Feedback in Higher Education. *J. Learn. Anal.* **2023**, *10*, 25–43. [CrossRef]
25. Annamalai, N.; Ab Rashid, R.; Hashmi, U.M.; Mohamed, M.; Alqaryouti, M.H.; Sadeq, A.E. Using chatbots for English language learning in higher education. *Comput. Educ. Artif. Intell.* **2024**, *5*, 100153. [CrossRef]
26. De Leon Evangelista, R. Artificial intelligence applications in competency-based assessment: A systematic review. *Int. J. Educ. Technol. High. Educ.* **2025**, *22*, 1–19.
27. Bond, M.; Buntins, K.; Bedenlier, S.; Zawacki-Richter, O.; Kerres, M. Mapping research in learning analytics and artificial intelligence in education: A systematic review. *Educ. Technol. Soc.* **2024**, *27*, 1–18.
28. Project. Report on the Perception of Teachers About the Future Impact of AI. 2025. Available online: https://www.paideiaproject.eu/wp-content/uploads/2025/06/Report-on-the-perception-of-teachers-about-the-future-impact-of-AI.pdf (accessed on 5 January 2026).

29. Luckin, R.; Holmes, W.; Griffiths, M.; Forcier, L.B. *Intelligence Unleashed: An Argument for AI in Education*; Pearson: London, UK, 2016.
30. Shute, V.J.; Rahimi, S. Review of computer-based assessment for learning in elementary and secondary education. *J. Comput. Assist. Learn.* **2017**, *33*, 1–19. [CrossRef]