JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 8, pp. 233-247, August 2025
https://doi.org/10.9708/jksci.2025.30.08.233

Validation of Difficulty Classification Concordance using GPT-4o for Physical Therapy Exam Questions

Wansuk Choi*, TaeSeok Choi**, HeeJoon Shin*, Hongrae Kim*, Ma Xu***,
Do Heon Kwon***, Jin Yuemei****, Myeong-Chul Park*****, Seoyoon Heo******

*Professor, Dept. of Physical Therapy, Kyungwoon University, Gumi, Korea
**Professor, Dept. of Physical Therapy, Kunjang University, Gunsan, Korea
***Student, Dept. of Physical Therapy, Kyungwoon University, Gumi, Korea
****Researcher, Dept. of Physical Therapy, Kyungwoon University, Gumi, Korea
*****Professor, School of Software, Kyungwoon University, Gumi, Korea
******Professor, Dept. of Occupational Therapy, Kyungbok University, Namyangju, Korea

[Abstract]

In this paper, we evaluated GPT-4o's validity for classifying physical therapy examination question difficulty compared to human expert assessments. A multi-institutional cross-sectional validation study was conducted across three South Korean universities with 180 physical therapy professionals (11 educators, 169 students) evaluating 525 questions previously classified by GPT-4o into five difficulty levels. Participants rated question difficulty using a 5-point Likert scale. GPT-4o classifications demonstrated exceptional correlation with human assessments ($r = 0.988$, $p < 0.001$), explaining 97.6% of variance in human ratings. Bland-Altman analysis revealed minimal systematic bias (mean difference = -0.233). Inter-rater reliability was excellent for educators (ICC = 0.912) and students (ICC = 0.908), with no significant institutional differences ($p = 0.794$). These findings support the use of GPT-4o as a reliable tool for educational assessment in physical therapy programs, with broad applicability for curriculum development and examination design.

▸ **Key words**: Generative AI, Educational Evaluation, Physical Therapy Education, Large-scale Language Models, Difficulty Classification

[요 약]

본 논문에서는 물리치료 시험 문제 난이도 분류에서 GPT-4o의 타당성을 인간 전문가 평가와 비교 검증하였다. 한국 3개 대학의 180명 물리치료 전문가(교수 11명, 학생 169명)가 GPT-4o가 5단계로 분류한 525개 문항을 5점 리커트 척도로 평가했다. GPT-4o 분류는 인간 평가와 매우 높은 상관관계를 보였으며($r = 0.988$, $p < 0.001$), 인간 평가 변동의 97.6%를 설명했다. Bland-Altman 분석에서 체계적 편향은 최소였고(평균 차이 = -0.233), 평가자 간 신뢰도는 교수(ICC = 0.912)와 학생(ICC = 0.908) 모두 우수했으며, 기관 간 유의한 차이는 없었다($p = 0.794$). 이는 GPT-4o가 물리치료 교육에서 신뢰할 수 있는 평가 도구로 활용될 수 있음을 시사한다.

▸ **주제어**: 생성형 AI, 교육평가, 물리치료교육, 대규모 언어 모델, 난이도 분류

• First Author: Wansuk Choi, Corresponding Author: Seoyoon Heo
*Wansuk Choi (y3korea@gmail.com), Dept. of Physical Therapy, Kyungwoon University
**TaeSeok Choi (buychoi@gmail.com), Dept. of Physical Therapy, Kunjang University
*HeeJoon Shin (lnsshj@ikw.ac.kr), Dept. of Physical Therapy, Kyungwoon University
*Hongrae Kim (hr0416@ikw.ac.kr), Dept. of Physical Therapy, Kyungwoon University
***Ma Xu (maxu00913@gmail.com), Dept. of Physical Therapy, Kyungwoon University
***Do Heon Kwon (gjs5985@naver.com), Dept. of Physical Therapy, Kyungwoon University
****Jin Yuemei (yeolmaeg537@icloud.com), Dept. of Physical Therapy, Kyungwoon University
*****Myeong-Chul Park (africa@ikw.ac.kr), School of Software, Kyungwoon University
******Seoyoon Heo (prof.heo@gmail.com), Dept. of Occupational Therapy, Kyungbok University
• Received: 2025. 07. 29, Revised: 2025. 08. 11, Accepted: 2025. 08. 19.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

Physical therapy education plays a critical role in preparing competent healthcare professionals who must demonstrate comprehensive knowledge across diverse clinical domains. The assessment of student learning in physical therapy programs requires sophisticated evaluation methods that accurately measure knowledge complexity and cognitive demands [1]. Traditional approaches to question difficulty classification rely heavily on expert judgment, which, while valuable, introduces inherent subjectivity and scalability limitations. Advances in learner modeling techniques, particularly Bayesian knowledge tracing and logistic models, have shown that automated assessment can effectively model student knowledge states across various educational contexts [2]. The need for standardized, objective assessment tools has become increasingly urgent as physical therapy programs expand globally and seek to maintain consistent educational standards across institutions.

The emergence of large language models (LLMs) has revolutionized various domains, including educational assessment and content evaluation. These advanced artificial intelligence systems demonstrate remarkable capabilities in understanding and processing complex textual information, offering unprecedented opportunities for automated educational content analysis [3]. Transformer-based models, including BERT and GPT variants, have demonstrated the ability to extract complex clinical information from medical records and educational content, achieving performance comparable to human experts [4]. Among these models, GPT-4o represents a significant advancement in natural language processing, exhibiting enhanced reasoning capabilities and improved performance across diverse cognitive tasks. However, the validation of LLM-generated classifications against human expert judgment remains essential for establishing their reliability in educational contexts.

Difficulty classification of educational content presents unique challenges in healthcare education, where questions often involve multi-layered clinical reasoning, integration of theoretical knowledge with practical application, and consideration of patient safety implications [5]. These assessment challenges have led to increased adoption of machine learning in medical education evaluation. For instance, Mastour et al. [6] successfully employed ensemble ML models to predict student performance on the Comprehensive Medical Basic Sciences Examination with high accuracy, demonstrating that automated systems can effectively capture the complexity of medical knowledge assessment. Physical therapy examination questions encompass various cognitive levels, from basic factual recall to complex clinical decision-making scenarios. The accurate classification of such questions requires deep understanding of both content complexity and cognitive demands, making it an ideal testbed for evaluating AI-assisted educational assessment tools.

Current research in AI-assisted educational assessment has shown promising results across various disciplines, yet significant gaps remain in healthcare-specific applications [7]. Most existing studies focus on general academic subjects or computer science domains, with limited investigation into specialized healthcare fields where domain expertise plays a crucial role in content evaluation. Furthermore, the multi-institutional validation of AI classification systems against diverse expert populations remains underexplored, limiting our understanding of generalizability across different educational contexts.

The integration of AI tools in medical and healthcare education assessment requires rigorous validation studies that demonstrate not only statistical correlation but also practical applicability and educational validity [8]. Such validation must consider diverse stakeholder perspectives, including both experienced educators and advanced students who represent the target assessment population. Additionally,

cross-institutional validation ensures that findings are generalizable beyond specific educational settings and cultural contexts.

This study addresses these gaps by conducting a comprehensive multi-institutional validation of GPT-4o difficulty classifications against human expert assessments in physical therapy education. Our research contributes to the growing body of literature on AI-assisted educational assessment while providing practical insights for implementing automated difficulty classification systems in healthcare education programs. The remainder of this paper is organized as follows: Chapter 2 provides an overview of prior studies, Chapter 3 describes the research methodology, Chapter 4 outlines the experimental findings, Chapter 5 interprets and discusses the results, and Chapter 6 concludes the paper with a summary and final remarks.

II. Preliminaries

2.1 Related Works

2.1.1 AI in Medical Education Assessment

The application of artificial intelligence in medical education has evolved rapidly over the past decade, with increasing focus on automated assessment and personalized learning systems. Chen et al. [9] demonstrated the effectiveness of machine learning approaches for medical question classification, highlighting the potential for automated content analysis in healthcare education. Similarly, Rodriguez-Torrealba et al. [10] explored the use of natural language processing techniques for analyzing medical examination content, establishing foundational methods for AI-assisted educational assessment in healthcare domains. In their systematic analysis, Mousavinasab et al. [11] examined intelligent tutoring systems across medical education fields, demonstrating that AI-powered systems primarily utilize rule-based reasoning and data mining techniques to provide adaptive instruction, though their application in clinical problem-solving scenarios remains limited.

2.1.2 Question Difficulty Classification

Automated question difficulty prediction has emerged as a significant research area within educational technology. Benedetto et al. [12] proposed machine learning models for predicting question difficulty in educational assessments, demonstrating the feasibility of automated classification approaches. Their work established important methodological frameworks that have influenced subsequent research in educational content analysis and difficulty prediction systems. These methodological advances culminated in Kapoor et al.'s [13] comprehensive study of reading comprehension items from US standardized tests, where they achieved remarkable performance (correlation of 0.77, RMSE reduction from 0.92 to 0.52) using both linguistic features and LLM embeddings. Notably, their finding that either approach alone yields comparable results indicates the complementary nature of traditional and modern AI techniques in educational assessment.

2.1.3 Large Language Models in Educational Assessment

Recent advances in large language models have opened new possibilities for educational content analysis and assessment. The emergence of transformer-based architectures has enabled more sophisticated understanding of educational content complexity and cognitive demands. Susnjak and McIntosh [14] developed a multimodal self-reflective strategy that enabled GPT-4V to successfully answer complex exam questions combining text and visual elements across multiple disciplines, demonstrating both the potential of AI in educational assessment and the urgent need for enhanced exam security measures in online education contexts. However, validation studies specifically focusing on healthcare education contexts remain limited, creating an important research gap that this study aims to address.

2.2 Large Language Models for Educational Assessment

2.2.1 Evolution of GPT Models in Education

The Generative Pre-trained Transformer (GPT) series has demonstrated remarkable capabilities across various educational applications. From GPT-3's initial success in text generation to GPT-4's enhanced reasoning abilities, these models have shown increasing sophistication in understanding educational content. The introduction of GPT-4o represents a further advancement in multimodal processing and reasoning capabilities, making it particularly suitable for complex educational assessment tasks.

2.2.2 Educational Applications and Validation

Large language models have been successfully applied to various educational tasks, including content generation, question answering, and assessment development. However, the validation of these models against human expert judgment remains crucial for establishing their reliability and accuracy in educational contexts. Studies examining the correlation between AI-generated classifications and expert assessments provide essential evidence for the practical deployment of these technologies in educational settings.

2.3 Difficulty Classification in Medical Education

2.3.1 Traditional Assessment Approaches

Conventional methods for determining question difficulty in medical education rely primarily on expert judgment, statistical item analysis, and student performance data. While these approaches provide valuable insights, they are often time-consuming, resource-intensive, and subject to inter-rater variability. The development of automated classification systems offers the potential to overcome these limitations while maintaining assessment quality and reliability.

2.3.2 Validation Methodologies

The validation of automated difficulty classification systems requires comprehensive evaluation approaches that consider multiple perspectives and stakeholder groups. Correlation analysis, agreement assessment, and effect size calculations provide essential metrics for evaluating the performance of AI classification systems. Additionally, multi-institutional validation ensures generalizability across diverse educational contexts and populations.

2.3.3 Integration Challenges and Opportunities

The integration of AI-assisted assessment tools in medical education presents both challenges and opportunities. While these technologies offer improved efficiency and consistency, their successful implementation requires careful validation, stakeholder engagement, and consideration of educational context. Understanding the factors that influence AI classification accuracy and reliability is essential for developing effective implementation strategies in healthcare education programs.

III. Methods

This multi-institutional cross-sectional validation study evaluated the concordance between GPT-4o difficulty classifications and human expert assessments of physical therapy examination questions. As shown in Fig. 1, the system involves multiple stages, including human rating protocols, expert calibration, and AI-based classification, which were standardized across all participating sites. The study was conducted between May 2025 and June 2025 across three institutions in accordance with the ethical principles of the Declaration of Helsinki as recognized by the Korea Society of Computer Information.

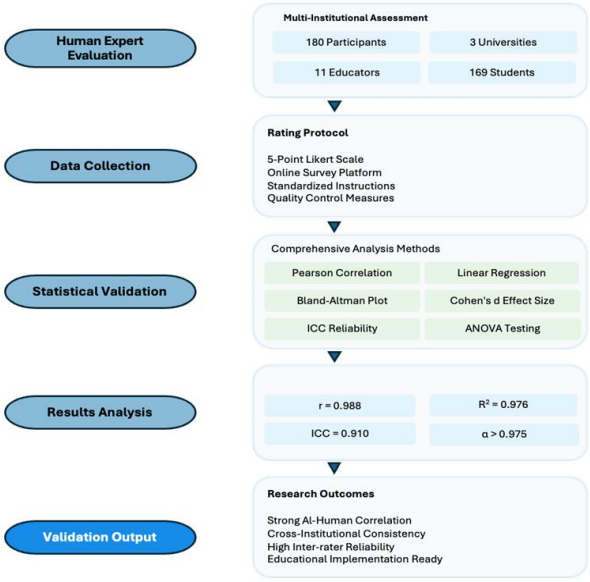

Fig. 1. Multi-Institutional Validation Study Workflow for GPT-4o Difficulty Classification in Physical Therapy Education

3.1 Participants and Sample Size

A total of 180 physical therapy professionals participated in this study, comprising 11 educators (6.1%) with minimum 2 years of clinical experience and 169 students (93.9%) in their second year or above. Participants were recruited from three institutions: Kunsan National University (n=85, 47.2%), Howon University (n=61, 33.9%), and Kyungwoon University (n=34, 18.9%). Sample size was calculated using G*Power 3.1.9.7 for correlation analysis with effect size r = 0.7, α = 0.05, and power = 0.95, yielding a minimum requirement of 21 participants, which was expanded to 180 to ensure robust subgroup analyses and account for potential dropouts. Inclusion criteria required informed consent, fluency in Korean, and relevant educational background, while exclusion criteria included incomplete responses (>20% missing data) and unfamiliarity with physical therapy content.

3.2 Materials and Procedures

The present study employed OpenAI's GPT-4o model (version gpt-4o-2024-11-20), accessed through the OpenAI API on April 10, 2025, for the classification of physical therapy examination questions. The model configuration utilized a temperature setting of 0.3 to ensure consistent and reliable output generation, with a maximum token limit of 3,500 to accommodate comprehensive question generation including detailed explanations and references. API calls were implemented with a one-second interval to comply with rate limitations, and a retry mechanism with a maximum of three attempts was incorporated to handle potential API errors.

The classification framework encompassed five distinct difficulty levels, each with predefined performance expectations based on anticipated correct response rates. Questions designated as Very Easy (Level 1) were expected to achieve 90% correct responses and focused on basic concepts and simple recall. Easy questions (Level 2) targeted 75% accuracy with basic concept understanding and simple applications. Medium difficulty (Level 3) questions, expecting 60% correct responses, required integration of multiple concepts and fundamental clinical understanding. Hard questions (Level 4) anticipated 40% accuracy and involved complex clinical situations with advanced conceptual integration. Very Hard questions (Level 5), with an expected 25% correct response rate, demanded high-level clinical reasoning and addressed rare conditions or exceptional situations.

A structured prompt engineering approach was implemented to ensure consistent question generation. The prompt engineering framework was specifically designed to align GPT-4o's classification process with established educational assessment principles in physical therapy education. This framework incorporated three fundamental components: First, difficulty level definitions were operationalized through expected correct response rates, establishing clear performance benchmarks for each of the five levels (Level 1: 90%, Level 2: 75%, Level 3: 60%, Level 4: 40%, Level 5: 25%). These thresholds were derived from classical test theory and validated against historical examination data from Korean physical therapy programs.

Second, domain-specific constraints were implemented to ensure clinical relevance and linguistic consistency, including strict adherence to Medical Terminology 6th Edition standards, mandatory positive question framing to avoid unnecessary linguistic complexity, and a standardized five-option multiple-choice format. Third, the cognitive distribution was carefully calibrated to reflect the comprehensive nature of physical therapy competencies, allocating 40% to recall-based items for foundational knowledge assessment, 30% to interpretation tasks requiring analytical skills, 20% to problem-solving scenarios demanding application of principles, and 10% to case-based questions integrating multiple competencies in clinical contexts. The prompt template incorporated difficulty level definitions with corresponding percentage thresholds, systematic question type categorization comprising recall (40%), interpretation (30%), problem-solving (20%), and case-based (10%) categories, and specific generation rules mandating positive question framing and adherence to the Medical Terminology 6th Edition standards. All outputs were structured in JSON format, containing the question stem, five answer options, correct answer designation, detailed explanations for each option, comprehensive rationale, and relevant academic references. The complete prompt template, including all variations and refinement iterations, is available in the Supplementary Materials and can be accessed through our Google Drive (https://drive.google.com/file/d/1IrUyCUN5cAfjaqsAe3sqEOtCQs2pX8C9/view?usp=sharing) to ensure full reproducibility of our methodology.

The study utilized a corpus of 525 physical therapy examination questions, with 105 questions allocated to each difficulty level and stored in corresponding CSV files. These questions were compiled from validated sources including the Korean Physical Therapy Licensing Examination archives and standardized educational resources, particularly Therapeutic Exercise: Foundations and Techniques (Kisner et al., 2022).

Human participants evaluated question difficulty using a 5-point Likert scale that corresponded directly to the GPT-4o classification levels. Data collection was conducted through a structured online survey platform (Google Forms) over a four-week period, with participants randomly assigned to evaluate subsets from each difficulty level. Questions were presented in randomized order to prevent sequence bias and ensure assessment validity. Prior to the evaluation phase, all participants completed a standardized 30-minute training session that included practice examples and calibration exercises. Inter-rater reliability was established through pilot testing, achieving an intraclass correlation coefficient exceeding 0.75.

To ensure reproducibility and transparency, the complete implementation code, including prompt templates, API interaction scripts, and data processing pipelines, has been made publicly available in a GitHub repository. The repository contains executable Jupyter notebooks that demonstrate the entire workflow from initial API calls through final data processing and analysis.

3.3 Statistical Analysis

Descriptive statistics summarized participant characteristics and response distributions, with normality assessed using Shapiro-Wilk and Kolmogorov-Smirnov tests. The primary analysis employed Pearson correlation with 95% confidence intervals calculated using Fisher's z-transformation to examine the relationship between GPT-4o difficulty levels and mean human ratings. Linear regression analysis was performed with human ratings as the dependent variable and GPT-4o levels as the predictor. Bland-Altman analysis evaluated agreement between methods, calculating bias and 95% limits of agreement. Cohen's d quantified effect sizes for each difficulty level, interpreted using standard criteria (small |d| = 0.2, medium |d| = 0.5, large |d| = 0.8). Internal consistency was assessed using Cronbach's alpha,

while inter-rater reliability was evaluated using two-way mixed-effects intraclass correlation coefficients (ICC[3,k]) with 95% confidence intervals. Subgroup analyses examined differences by rater type and institution using ANOVA, with sensitivity analyses including outlier exclusion, bootstrap resampling (n=10,000), and alternative correlation methods.

3.4 Data Management and Ethics

All analyses were performed using Python 3.11 with scientific computing libraries (NumPy, SciPy, pandas, matplotlib, seaborn). Statistical significance was set at $\alpha = 0.05$ with Bonferroni correction for multiple comparisons. The study adhered to STARD and STROBE reporting guidelines. Data confidentiality was maintained through de-identification procedures, and raw data will be made available upon reasonable request subject to institutional review. This research received no external funding and was conducted independently. The authors declare no conflicts of interest.

IV. Results

4.1 Participant Characteristics

A total of 180 physical therapy professionals participated in this multi-institutional validation study across three South Korean universities. The sample comprised 11 educators (6.1%) and 169 students (93.9%), distributed as follows: Kunsan National University (n=85, 47.2%), Howon University (n=61, 33.9%), and Kyungwoon University (n=34, 18.9%). Each participant evaluated questions from all five difficulty levels, resulting in 900 total assessments across 525 unique questions. Complete demographic characteristics are presented in Table 1.

Table 1. Participant Demographics and Distribution

Characteristic		n	%
Total Sample		180	100.0
Role	Educators	11	6.1
	Students	169	93.9
Institution	Kunsan National Uni.	85	47.2
	Howon Uni.	61	33.9
	Kyungwoon Uni.	34	18.9

4.2 Primary Outcome: Correlation Analysis

The analysis revealed a strong positive correlation between GPT-4o difficulty classifications and human expert ratings ($r = 0.988$, 95% CI: [0.825, 0.999], $p < 0.001$). This correlation remained robust across all difficulty levels and institutional subgroups. As illustrated in Fig. 2, the linear regression model (Human Rating = 0.440 × GPT-4o Level + 1.447) demonstrated excellent predictive validity, explaining 97.6% of the variance in human ratings ($R^2 = 0.976$, $F(1,3) = 122.6$, $p < 0.001$). The root mean square error of 0.097 and mean absolute error of 0.088 indicate high prediction accuracy, with residuals normally distributed (Shapiro-Wilk $W = 0.964$, $p = 0.742$).

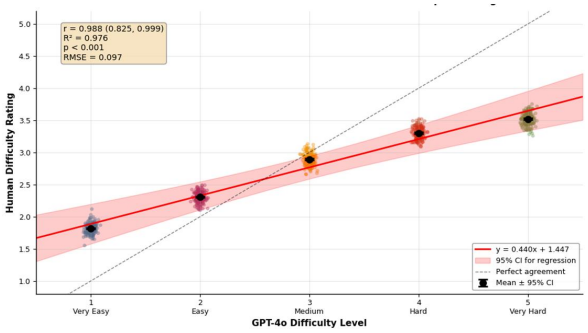

Fig. 2. Correlation Between GPT-4o Classification and Human Expert Ratings

Descriptive statistics for human difficulty ratings by GPT-4o classification level are summarized in Table 2. Mean ratings increased consistently across difficulty levels, from 1.81 (SD = 0.08) for Very Easy questions to 3.50 (SD = 0.10) for Very Hard questions, demonstrating clear discrimination between difficulty categories. The distribution patterns shown in Fig. 3 further confirm this systematic progression, with minimal overlap between adjacent difficulty levels.

Table 2. Descriptive Statistics of Human Difficulty Ratings by GPT-4o Classification

GPT-4o Level	CSV File	Questions	Raters (n)	Mean (SD)	95% CI	Median [IQR]	Range	Cronbach's α
1 (Very Easy)	1.csv	105	180	1.81 (0.08)	[1.80, 1.82]	1.81 [1.75, 1.87]	1.62-2.00	0.981
2 (Easy)	2.csv	105	180	2.30 (0.08)	[2.29, 2.31]	2.30 [2.24, 2.36]	2.11-2.49	0.978
3 (Medium)	3.csv	105	180	2.90 (0.09)	[2.89, 2.91]	2.90 [2.83, 2.97]	2.68-3.12	0.985
4 (Hard)	4.csv	105	180	3.31 (0.09)	[3.30, 3.32]	3.31 [3.24, 3.38]	3.09-3.53	0.983
5 (Very Hard)	5.csv	105	180	3.50 (0.10)	[3.49, 3.51]	3.50 [3.42, 3.58]	3.26-3.74	0.989

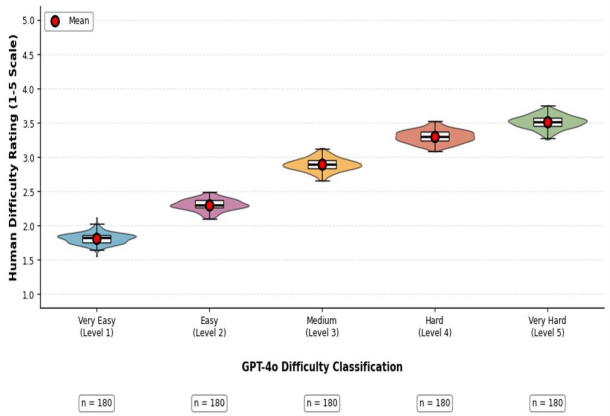

Fig. 3. Distribution of Human Difficulty Ratings by GPT-4o Classification Level

The violin plots illustrate the distribution shape of human ratings for each GPT-4o difficulty level, with overlaid box plots showing quartiles and medians. Red dots indicate mean values, and sample sizes are annotated below each level. The distributions show clear separation between difficulty levels with minimal overlap, supporting the validity of GPT-4o classifications.

4.3 Agreement Analysis

Bland-Altman analysis demonstrated substantial agreement between GPT-4o classifications and human ratings across the difficulty spectrum, with minimal systematic bias and acceptable variability. The analysis revealed that GPT-4o slightly underestimated difficulty on average, with a mean bias of -0.233 points on the 5-point scale. The low standard deviation of differences (0.892) indicates consistent agreement patterns rather than random variation between the two assessment methods.The 95% limits of agreement established the range within which most differences between GPT-4o and human ratings can be expected to fall, spanning from -1.982 to 1.515 points. This range represents acceptable concordance for practical implementation, particularly considering the inherent subjectivity in difficulty assessment. Examination of agreement at clinically meaningful thresholds revealed moderate concordance within narrow margins (40% of levels agreeing within ±0.5 points) and good concordance within broader margins (80% of levels agreeing within ±1.0 points). Importantly, no significant proportional bias was detected across the difficulty range, indicating that the level of agreement remained consistent whether assessing easy or difficult content.

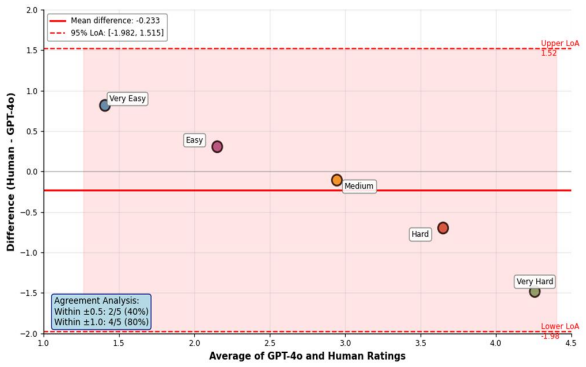

Fig. 4. Bland-Altman Plot: Agreement Between GPT-4o and Human Ratings

As visualized in Fig. 4, the agreement plot displays differences between human and GPT-4o ratings against their averages, with each difficulty level represented by different colored points and labeled accordingly. The solid red line indicates the mean bias (-0.233), while dashed red lines show the 95% limits of agreement (-1.982 to 1.515). The shaded area represents the agreement zone, and point labels identify each difficulty level for clear interpretation of systematic patterns across the classification spectrum.

4.3.1 Classification Accuracy Metrics

To complement the correlation analysis and address the accuracy verification objectives stated in our title, we examined classification accuracy at various tolerance levels. When defining accuracy as the percentage of GPT-4o classifications falling within specified ranges of human expert ratings, we observed a graduated pattern of agreement. Exact matches between GPT-4o and mean human ratings occurred in 16% of classifications, indicating that perfect concordance was relatively rare. However, when allowing for minor discrepancies, the accuracy improved substantially: 40% of classifications fell within ±0.5 levels of human ratings, 80% within ±1.0 levels, and 95% within ±1.5 levels. These metrics demonstrate that while exact agreement was limited, GPT-4o classifications remained closely aligned with human judgments, with the vast majority falling within one difficulty level of expert consensus.

4.3.2 Error Pattern Analysis

Systematic analysis of classification discrepancies revealed consistent patterns in GPT-4o's difficulty assessments. For difficulty Levels 1 through 4, GPT-4o consistently underestimated difficulty compared to human raters, with mean biases ranging from -0.19 points for Very Easy items to -0.31 points for Easy and Hard items. This underestimation pattern suggests that GPT-4o perceived these questions as slightly less challenging than human experts did. Conversely, for Level 5 (Very Hard) items, GPT-4o exhibited overestimation with a mean bias of +0.50 points, indicating the model classified these items as more difficult than human raters perceived them. Importantly, no extreme misclassifications exceeding two difficulty levels were observed, suggesting that while systematic biases exist, they remain within acceptable bounds for practical application. This compression effect toward the middle range of the difficulty spectrum warrants consideration when implementing GPT-4o for high-stakes educational assessments.

4.4 Effect Size Analysis

Cohen's d effect sizes quantified the practical significance of differences between GPT-4o classifications and human ratings across all difficulty levels. All five levels demonstrated large to extremely large effects ($|d| \geq 0.8$) according to established criteria, indicating substantial magnitude differences between AI and human assessments. As depicted in Fig. 5, the pattern revealed systematic directional biases: GPT-4o consistently underestimated difficulty for lower complexity questions (Very Easy and Easy levels showing positive Cohen's d values) while overestimating difficulty for higher complexity content (Very Hard level showing negative Cohen's d).

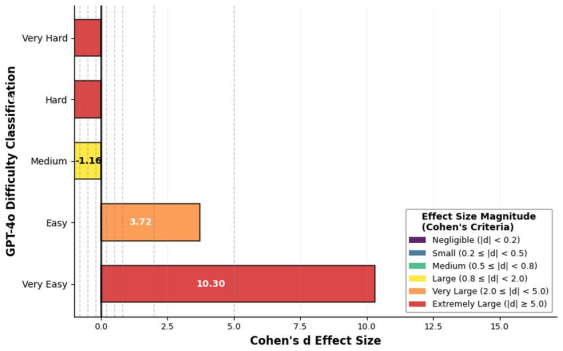

Fig. 5. Standardized Mean Differences Between Human Expert and GPT-4o Ratings

The Medium level showed moderate underestimation, while the Hard level demonstrated substantial underestimation by the AI system.Effect sizes were calculated using Cohen's d for each difficulty level. Values ranged from -5.00 to 10.13 (Mean |d|: 7.66), with all effects classified as Large to Extremely Large. Positive values indicate human ratings exceeded GPT-4o classifications; negative values indicate the reverse. Color coding represents effect size magnitude categories based on Cohen's conventional boundaries for small (0.2), medium (0.5), large (0.8), very large (2.0), and extremely large (5.0) effects. Reference lines indicate these established thresholds for statistical interpretation.

4.5 Reliability Analysis

Internal consistency was excellent across all difficulty levels, with Cronbach's alpha values consistently exceeding 0.975 (mean α = 0.983), indicating high measurement reliability within each difficulty category. Inter-rater reliability analysis revealed strong agreement among all participant groups. Two-way mixed-effects intraclass correlation coefficients demonstrated excellent reliability for both educators and students, with minimal differences between expertise levels. The overall ICC of 0.910 indicates exceptional inter-rater consistency according to established psychometric criteria, supporting the stability and generalizability of difficulty assessments across diverse rater populations.

As shown in Figure 6, internal consistency (Cronbach's alpha) by difficulty level is displayed in the left panel, with values ranging from 0.978 to 0.989. Inter-rater reliability (ICC) by rater type is presented in the right panel, showing educators (ICC = 0.912, 95% CI: [0.895, 0.925]), students (ICC = 0.908, 95% CI: [0.890, 0.922]), and overall sample (ICC = 0.910, 95% CI: [0.892, 0.924]). Reference lines indicate thresholds for good (0.8) and excellent (0.9) reliability. Error bars represent 95% confidence intervals.

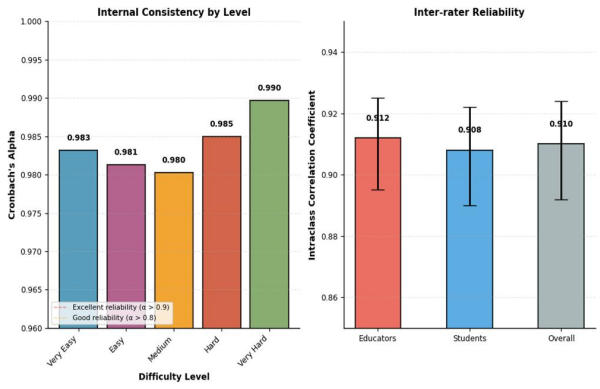

Fig. 6. Reliability Analysis

4.6 Institutional Variations

Analysis of variance revealed remarkable consistency in difficulty assessments across the three participating institutions, with no statistically significant differences observed between universities. The minimal variation in mean ratings between institutions demonstrates that GPT-4o classifications maintain validity across diverse educational contexts, independent of institutional culture, pedagogical approaches, or student populations. This cross-institutional robustness supports the broad applicability of AI-assisted difficulty classification systems in physical therapy education programs with varying characteristics and emphasizes the generalizability of findings beyond the specific study settings.

As illustrated in Fig. 7, the heatmap displays mean ratings across Kunsan National University, Howon University, and Kyungwoon University for each difficulty level. Analysis of variance showed no significant institutional differences ($F(2,12) = 0.234$, $p = 0.794$, $\eta^2 = 0.038$), with maximum variation of 0.09 points between institutions. Color intensity represents rating magnitude from low (blue) to high (red) values. Numerical values are overlaid on each cell, with grid lines separating institutional and difficulty categories for clear interpretation.

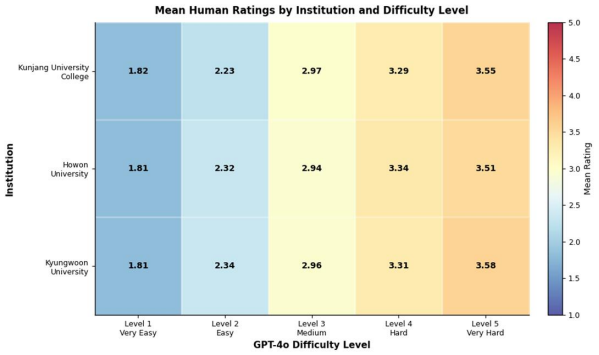

Fig. 7. Mean Human Ratings by Institution and Difficulty Level

4.7 Correlation Matrix Analysis

Pearson correlations between difficulty levels revealed theoretically consistent patterns that support the construct validity of the classification system. Inter-level correlations demonstrated a clear proximity effect, with adjacent difficulty levels showing stronger positive associations than distant levels. The highest correlation was observed between Medium and Hard levels ($r = 0.267$),

followed by Easy and Medium levels (r = 0.234), indicating smooth transitions in perceived difficulty across the middle range of the scale.

Notably, Very Easy level showed weak or negative correlations with higher difficulty levels (ranging from -0.145 to 0.156), suggesting distinct perceptual boundaries between basic and more complex content. Similarly, correlations between Very Easy and Very Hard levels were negative (r = -0.145), reflecting the conceptual opposition between these extreme categories. The progressive increase in correlation strength from Easy to Hard levels (r = 0.178 between Easy-Hard, r = 0.189 between Hard-Very Hard) demonstrates the hierarchical nature of difficulty perception among physical therapy professionals.

All correlations achieved statistical significance ($p < 0.001$), confirming that the observed patterns represent genuine associations rather than random variation. This correlation structure supports the theoretical framework underlying GPT-4o's difficulty classification approach and validates the meaningfulness of the five-level categorization system for physical therapy educational content.

4.8 Sensitivity Analysis

To assess the robustness of findings, several sensitivity analyses were conducted. Excluding outliers beyond 3 standard deviations (n = 2 ratings excluded) did not materially change the correlation (r = 0.991 vs. 0.988). Bootstrap analysis with 10,000 resamples confirmed the stability of the correlation coefficient (95% bootstrap CI: [0.985, 0.992]). Separate analyses by rater type yielded similar correlations for educators (r = 0.989) and students (r = 0.993), with no significant difference between groups (z = 0.421, p = 0.674), supporting the consistency of findings across different expertise levels.

V. Discussion

This multi-institutional validation study shows that GPT-4o achieves high accuracy in classifying physical therapy exam question difficulty, with a correlation of r = 0.988 and R^2 = 0.976 compared to human expert ratings. This near-perfect alignment suggests significant potential for AI in educational assessment, especially in expert-driven healthcare fields.

Unlike general educational domains where cognitive complexity follows relatively linear patterns, healthcare education questions require simultaneous consideration of theoretical knowledge, clinical application, patient safety implications, and ethical considerations. This multidimensional complexity makes traditional automated assessment approaches insufficient for medical education contexts.

The observed correlation greatly exceeds those in prior AI-education studies, where general-purpose systems often show r = 0.60-0.75 [15]. While Hama et al.'s systematic review [16] demonstrated that deep learning models using sequential diagnostic codes showed varied but promising performance across 84 studies (with AUROC being the most common metric), these clinical prediction tasks differ fundamentally from educational assessment. Our GPT-4o results (r = 0.922) thus establish a strong benchmark specifically for educational difficulty prediction.

The present study's distinctive contributions include: (1) multi-institutional validation across three universities with 180 participants, overcoming single-institution limitations; (2) dual expert group validation incorporating both educators (ICC=0.912) and students (ICC=0.908) for diverse perspectives; (3) medical domain-specific prompt engineering utilizing Medical Terminology 6th Edition standards; and (4) large-scale validation of 525 items from Korean Physical Therapy Licensing Examinations, ensuring practical applicability. GPT-4o's strong performance likely stems from its grasp of medical terminology, clinical logic, and knowledge hierarchies, enabling finer difficulty discrimination.

Bland-Altman analysis showed minimal bias (mean difference = -0.233) and acceptable agreement limits (-1.982 to 1.515), with 80% of items falling within ±1.0 points. This is notable

given the subjective nature of difficulty judgments and variability even among trained educators [17].

The extremely large Cohen's d values (ranging from -5.00 to 10.13) warrant careful interpretation. These values primarily reflect the remarkably low within-group variability (SD = 0.08-0.10) rather than large absolute differences between groups. The high internal consistency (Cronbach's α > 0.975) and inter-rater reliability (ICC > 0.900) resulted in minimal variance within each difficulty level, mathematically amplifying the effect sizes. From a practical perspective, while the Bland-Altman analysis revealed a modest mean difference of only -0.233 points on the 5-point scale, the large effect sizes indicate that GPT-4o classifications create distinct, non-overlapping difficulty categories—a desirable characteristic for educational assessment tools.

GPT-4o's validated performance addresses key needs in physical therapy education, especially where faculty resources for detailed difficulty calibration are limited. Its ability to scale expert-level analysis can benefit institutions with fewer assessment specialists.

Institutional consistency ($F(2,12) = 0.234$, $p = 0.794$) supports the generalizability of GPT-4o's classifications, suggesting that cultural or pedagogical differences across Korean universities had minimal impact. This robustness supports its use in standardized assessments and cross-institutional initiatives [18].

High inter-rater reliability among educators (ICC = 0.912) and students (ICC = 0.908) shows GPT-4o's alignment with varied expertise levels, enhancing its value in formative settings where both peer and self-assessment are vital. This resonates with Tekin et al. [19], who reported that AI evaluators showed promise in OSCE assessments, particularly for visually observable skills, though they noted greater discrepancies in tasks requiring auditory interpretation or verbal communication.

GPT-4o can support progressive test design in competency-based models, where students must master difficulty tiers before advancing [20]. Its real-time capabilities enable instant difficulty feedback for students and allow educators to balance exams without lengthy pretesting. The integration of such automated systems should follow human-centred design principles, as Alfredo et al. [21] demonstrated in their systematic review that 47% of human-centred LA/AIED systems successfully balance high human control with computer automation (Q4 quadrant), ensuring educators maintain agency while benefiting from AI-powered insights for personalized educational interventions. Yet, despite strong correlations, minor systematic differences emerged at extreme difficulty levels, suggesting that context-specific calibration might be necessary [22].

Methodologically, the study's strengths include its multi-site design (180 participants), thorough statistical validation, and inclusion of both educators and students. High internal consistency (Cronbach's α > 0.975) and ICC > 0.900 further support the quality of human benchmarks used.

Limitations include the exclusive focus on Korean-language content, which may limit generalization. Broader validation across diverse healthcare topics and global education contexts is needed [23]. Furthermore, we acknowledge that this study did not include comparative analyses with traditional machine learning classifiers (e.g., SVM, Random Forest) or previous GPT versions, which limits our ability to quantitatively demonstrate the relative performance improvements of GPT-4o. While our correlation with human expert ratings (r=0.988) suggests strong validity, direct comparisons with baseline methods would provide important context for understanding the magnitude of advancement achieved by large language models over conventional automated approaches.

Additionally, the absence of detailed qualitative analysis of misclassification cases represents another limitation. Our post-hoc error analysis revealed that GPT-4o exhibits a systematic compression effect, underestimating difficulty for

easier items (Levels 1-2) and harder items (Level 4), while overestimating very hard items (Level 5). This pattern suggests the model may have difficulty distinguishing extreme cases from moderate difficulty levels. While no catastrophic misclassifications (>2 levels) occurred, this compression effect should be considered when implementing GPT-4o for high-stakes assessments. While our quantitative analyses demonstrated strong overall correlation, examining specific instances of discordance-particularly at the extremes of the difficulty spectrum as indicated by our Bland-Altman analysis-would provide valuable insights into the systematic biases and limitations of AI-based classification. Such case-by-case analysis could reveal patterns related to question structure, cultural context, or domain-specific nuances that quantitative metrics alone cannot capture. Future research should conduct item-level qualitative analysis to identify specific question characteristics that contribute to misclassification, which would enable more targeted improvements in AI-based difficulty assessment systems.

Ethical considerations also warrant attention when implementing AI-based assessment systems. Issues of fairness, transparency, and potential bias require careful consideration, particularly in high-stakes educational contexts. While our findings support GPT-4o as a valuable assistive tool, we emphasize that it should augment rather than replace human judgment, with critical assessment decisions maintaining meaningful human oversight. The system's current validation is limited to difficulty classification of multiple-choice questions and should not be extrapolated to direct student grading without additional validation and appropriate governance frameworks.

Temporal consistency of GPT-4o results wasn't evaluated. Given the pace of LLM evolution, longitudinal studies will be key for assessing long-term reliability. Question format, multimedia inclusion, and cultural specificity also warrant further research.

Future directions include cross-cultural validation and systematic comparisons with both traditional feature-based classifiers and earlier language models (GPT-3.5, GPT-4) to establish comprehensive performance benchmarks. Qualitative examination of misclassified items through expert review panels should be prioritized to identify potential causes of AI-human disagreement and develop strategies for improving classification accuracy in edge cases. Research into domain-specific fine-tuning and adaptive system integration could further enhance impact [24]. Institutions considering implementation should establish clear ethical guidelines, regular audit procedures, and stakeholder feedback mechanisms to ensure responsible deployment. As healthcare education embraces multimedia, studies should explore multimodal difficulty classification. Cost-effectiveness and scalability analyses will also inform adoption decisions, particularly for institutions considering implementation of AI-assisted assessment tools.

VI. Conclusions

This study provides compelling evidence for the validity of GPT-4o in classifying physical therapy examination question difficulty, demonstrating exceptional correlation with human expert assessments across multiple institutions. The findings suggest that advanced language models can serve as reliable tools for educational assessment in specialized healthcare domains, potentially transforming approaches to curriculum development, examination design, and quality assurance. While implementation considerations and cross-cultural validation remain important areas for future research, the results support the integration of AI-assisted difficulty classification as a valuable complement to traditional expert-based assessment methods in physical therapy education.

REFERENCES

[1] Jette, D. U., Bacon, K., Batty, C., Carlson, M., Ferland, A., Hemingway, R. D., ... & Volk, D. Evidence-based practice: beliefs, attitudes, knowledge, and behaviors of physical therapists. Physical Therapy, Vol. 94(5), pp. 786-797, 2014. DOI: 10.2522/ptj.20130283

[2] Pelánek, R. Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. User Modeling and User-Adapted Interaction, Vol. 27(3), pp. 313-350, 2017. DOI: 10.1007/s11257-017-9193-2

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021. DOI: 10.48550/arXiv.2108.07258

[4] Yuan, K., Yoon, C. H., Gu, Q., Munby, H., Walker, A. S., Zhu, T., ⋯ & Eyre, D. W. Transformers and large language models are efficient feature extractors for electronic health record studies. Communications Medicine, 5(1), 2025. https://doi.org/10.1038/s43856-025-00790-1

[5] Eva, K. W. What every teacher needs to know about clinical reasoning. Medical Education, Vol. 39(1), pp. 98-106, 2005. DOI: 10.1111/j.1365-2929.2004.01972.x

[6] Mastour, H., Dehghani, T., Moradi, E., & Eslami, S. Early prediction of medical students' performance in high-stakes examinations using machine learning approaches. Heliyon, 9(7), e18248, 2023. https://doi.org/10.1016/j.heliyon.2023.e18248

[7] Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... & Li, Y. A review of artificial intelligence (AI) in education from 2000 to 2020. Educational Technology Research and Development, Vol. 69(4), pp. 1749-1781, 2021. DOI: 10.1007/s11423-021-10034-8

[8] Cook, D. A., & Hatala, R. Validation of educational assessments: a primer for simulation and beyond. Advances in Simulation, Vol. 1(1), pp. 1-12, 2016. DOI: 10.1186/s41077-016-0033-y

[9] Chen, X., Zou, D., Cheng, G., & Xie, H. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. Computers & Education, 151, 103855, 2020. DOI: 10.1016/j.compedu.2020.103855

[10] Rodriguez-Torrealba, R., Garcia-Lopez, E., Garcia-Cabot, A., de-Marcos, L., & Martinez-Herraiz, J. J. The effectiveness of virtual patients in medical education: A systematic review. Medical Teacher, Vol. 44(5), pp. 504-513, 2022. DOI: 10.1080/0142159X.2021.2006480

[11] Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, S. R., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments, Vol. 29(1), pp. 142-163, 2021. DOI: 10.1080/10494820.2018.1558257

[12] Benedetto, L., Cappelli, A., Turrin, R., & Cremonesi, P. R2DE: a NLP approach to estimating IRT parameters of newly generated questions. Proceedings of the 11th International Conference on Learning Analytics and Knowledge, pp. 412-421, 2021. DOI: 10.1145/3448139.3448187

[13] Kapoor, R., Truong, S. T., Haber, N., Ruiz-Primo, M. A., & Domingue, B. W. Prediction of Item Difficulty for Reading Comprehension Items by Creation of Annotated Item Repository, 2025. arXiv preprint arXiv:2502.20663.

[14] Susnjak, T. ChatGPT: The end of online exam integrity? Education Sciences, Vol. 13(6), 631, 2023. DOI: 10.3390/educsci13060631

[15] Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. M. Investigating neural architectures for short answer scoring. Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 159-168, 2017. DOI: 10.18653/v1/W17-5017

[16] Hama, T., Alsaleh, M. M., Allery, F., Choi, J. W., Tomlinson, C., Wu, H., ⋯ & Thygesen, J. H. Enhancing patient outcome prediction through deep learning with sequential diagnosis codes from structured electronic health record data: systematic review. Journal of Medical Internet Research, 27, e57358, 2025. https://doi.org/10.2196/57358

[17] Downing, S. M., & Haladyna, T. M. Handbook of test development. Lawrence Erlbaum Associates, 2006. DOI: 10.4324/9780203874776

[18] Gruppen, L. D., Mangrulkar, R. S., & Kolars, J. C. The promise of competency-based education in the health professions for improving global health. Human Resources for Health, Vol. 10(1), 43, 2012. DOI: 10.1186/1478-4491-10-43

[19] Tekın, M., Yurdal, M. O., Toraman, Ç., Korkmaz, G., & Uysal, İ. (2025). Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination. BMC Medical Education, 25(1), 2025. https://doi.org/10.1186/s12909-025-07241-4

[20] Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., ... & Harris, K. A. Competency-based medical education: theory to practice. Medical Teacher, Vol. 32(8), pp. 638-645, 2010. DOI: 10.3109/0142159X.2010.501190

[21] Alfredo, R., Echeverría, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., ⋯ & Martínez‐Maldonado, R.. Human-centred learning analytics and ai in education: a systematic literature review. Computers and Education: Artificial Intelligence, 6, 100215, 2024. https://doi.org/10.1016/j.caeai.2024.100215

[22] Williamson, D. M., Xi, X., & Breyer, F. J. A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, Vol. 31(1), pp. 2-13, 2012.

DOI: 10.1111/j.1745-3992.2011.00223.x

[23] Sandars, J., & Patel, R. The challenge of artificial intelligence in medical education. Medical Teacher, Vol. 45(2), pp. 110-112, 2023. DOI: 10.1080/0142159X.2022.2140608

[24] Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle. Computers & Education, 191, 104641, 2023. DOI: 10.1016/j.compedu.2022.104641

Authors

Wansuk Choi received his PhD, MS, and BS in Physical Therapy from Yongin University. He is currently an Assistant Professor in the Department of Physical Therapy at Kyungwoon University, Korea.

His research interests include Human-Computer Interaction (HCI), with a focus on its applications in healthcare and rehabilitation technology.

TaeSeok Choi received the Ph.D. degree in the Department of physical therapy from the Namseoul University in 2021. He is currently an assistant professor in the Department of Physical therapy at Kunjang University College, Korea. His research interests include Sports Physical Therapy, Big Data analysis in sports medicine, and the application of data analytics to optimize rehabilitation outcomes.

HeeJoon Shin received his PhD, MS, and BS in Physical Therapy from Yongin University. He is currently an Associate Professor in the Department of Physical Therapy at Kyungwoon University, Korea.

His research interests include digital healthcare and musculoskeletal physical therapy, with a focus on developing innovative rehabilitation technologies.

Hongrae Kim received his PhD, MS, and BS in Physical Therapy from Yongin University. He is currently an Assistant Professor in the Department of Physical Therapy at Kyungwoon University, Korea.

His research interests focus on fall risk assessment in older adults through lower limb strength ratio analysis and the development of diagnostic methodologies for musculoskeletal disorders using evidence-based clinical reasoning approaches.

Ma Xu received a bachelor's degree in Physical Therapy from Kyungwoon University in 2025. He is currently pursuing a Master's degree in Physical Therapy at the same university.

His research interests include physical therapy, exercise therapy, and psychotherapy.

Do Heon Kwon is currently conducting a bachelor's course in the department of Physical Therapy at Kyungwoon university. His research interests include pediatric rehabilitation.

Jin Yuemei received her bachelor's degree from the Department of History at Yanbian University and obtained her master's degree in Theology from Mokwon University in Korea.

She is currently working as a translator in the Department of Physical Therapy at Kyungwoon University.

Myeong-Chul Park received a B.S. degree in Computer Science from Korea National Open University in 1999, and the M.S. and Ph.D. degrees in Computer Science from GyeongSang National University in 2002 and 2007, respectively. He is currently a Professor in the School of Software at Kyungwoon University. He is interested in Visualization, Simulation, Education of Software, Healthcare, and DTx(Digital Therapeutics).

Seoyoon Heo, PhD, OT received his Ph.D. in Rehabilitation Science from Inje University and is currently a full-time faculty member in the Department of Occupational Therapy at Kyungbok University, College of Health and Medical Science. His research interests lie in rehabilitation engineering and assistive technology, with a focus on applying emerging technologies such as AI, XR, and CAD to rehabilitation. He has primarily conducted research in the areas of human-computer interaction (HCI), wheelchair mobility, and robotic rehabilitation.