International Journal of Public Health and Medical Research

ISSN: 3006-7707 (Print), ISSN: 3006-3752 (Online) | Volume 3, Number 3, Year 2025
DOI: https://doi.org/10.62051/ijphmr.v3n3.02
Journal homepage: https://wepub.org/index.php/IJPHMR/index

Development and Reliability-Validity Testing of a Knowledge, Attitude, and Behavior (KAB)-Based Scale for University Students' Academic Use of AI Tools

Yangli Ou [1, 2], Nancy A. Aguila [1], Lulu Xiao [2], Zhiyuan Zhang [3], Fangzhu Ouyang [4, *]

[1] College of Nursing and Allied Health Sciences, St. Paul University Manila, Manila, Philippines

[2] College of Nursing, South China University, Hengyang, China

[3] The Second Affiliated Hospital, South China University, Hengyang, China

[4] The Affiliated Nanhua Hospital, South China University, Hengyang, China

*Corresponding Author: Fangzhu Ouyang

ABSTRACT

To develop and validate a scale for assessing the knowledge and behaviors related to academic use norms of AI tools among university students, focusing on reliability and validity. Grounded in the Knowledge-Attitude-Behavior (KAB) theory, a pool of scale items was constructed through a review of the literature, semi-structured interviews, and Delphi expert consultations. A pilot test was conducted with 140 undergraduate nursing students from a certain university in China, employing item analysis and exploratory factor analysis to refine the scale items. A formal assessment was then conducted with 220 nursing undergraduates who had used generative AI tools for at least three months. Reliability analysis and confirmatory factor analysis were performed using SPSS and AMOS. The scale consists of 2 subscales, 3 dimensions, and 14 items. Exploratory factor analysis extracted three common factors, with a cumulative variance explanation of 70.878%. The Cronbach's α coefficients for the subscales were 0.925 and 0.855, respectively, while the split-half reliability coefficients were 0.891 and 0.860. The fit indices for the confirmatory factor model were as follows: $\chi^2/df=2.978$, GFI=0.859, PGFI=0.606, RMSEA=0.098, RMR=0.048, CFI=0.920, and PCFI=0.748. The scale demonstrates strong reliability and validity, making it suitable for assessing university students' knowledge and behaviors regarding the academic use of AI tools.

KEYWORDS

College students; Generative AI tools; Academic misconduct; KAB theory

1. INTRODUCTION

Academic integrity refers to the principle that educators, students, researchers, and all members of the academic community should conduct themselves with honesty, trust, fairness, respect, and responsibility in academic activities [1]. Violations of academic integrity are commonly referred to as academic misconduct or dishonesty. However, advancements in information technology have made it easier for university students to access information resources, potentially facilitating academic misconduct such as plagiarism, copying, and cheating [2].

In recent years, generative artificial intelligence (AI) models have rapidly evolved, demonstrating remarkable capabilities in answering questions, summarizing texts, and engaging in open-domain conversations [3]. Students can leverage AI tools to conveniently access vast amounts of information for completing assignments, essays, reports, and academic writing tasks [4, 5]. According to a survey conducted in January 2023 by the online education platform Study.com, which polled 1,000 U.S. college students aged 18 and above, more than 90% of respondents were aware of ChatGPT, with over 89% having used it to complete assignments, 48% utilizing it for quizzes, 53% for writing essays, and 22% for generating essay outlines [5]. Educators have expressed concerns that the use of AI tools by students may increase the risks of plagiarism and cheating, thereby undermining academic integrity. Furthermore, they worry that excessive reliance on AI could negatively impact students' creativity and critical thinking skills. As a result, there is a growing call for the regulation of students' use of AI tools in academic settings [6].

In response to this issue, China's Ministry of Foreign Affairs released the Position Paper on Strengthening Ethical Governance of Artificial Intelligence in November 2022, advocating the principle of "ethics first" [7]. In July 2024, Chinese scholar Junfeng Yang [8] and colleagues established ethical norm indicators at the school level from a macro perspective on the application of AI in education. However, from a micro perspective, assessment tools for evaluating university students' knowledge of AI tool usage norms and their tendencies toward academic misconduct remain insufficient.

Based on this, the present study develops a scale to assess university students' knowledge of AI tool usage norms and their academic behaviors, grounded in the Knowledge-Attitude-Behavior (KAB) theory, also known as the Technology Acceptance Model. The study aims to evaluate students' cognitive understanding of AI tool usage norms and identify potential academic misconduct when using AI tools, providing a foundation for optimizing institutional policies and implementing academic integrity training in higher education. KAB theory, which comprises three key components—knowledge, attitude, and behavior—has been widely applied in scale development, training, and intervention design [9-11], making it well-suited for the development of this assessment scale.

2. METHODS

2.1. Establishment of the Research Team

The research team consists of five members, including three university faculty members specializing in nursing education research, one clinician with expertise in AI tools and clinical teaching, and one clinical nurse. The primary responsibilities of the team include the following: conducting a literature review to establish the theoretical foundation and content framework of the scale, developing a semi-structured interview outline, conducting interviews and summarizing the results, selecting and confirming the Delphi expert panel members, designing the expert consultation questionnaire, compiling and analyzing expert feedback, implementing the formal survey, and evaluating the reliability and validity of the questionnaire.

2.2. Constructing a Framework for College Students' Perceived Usefulness and Usage Behavior Scale of AI Tools

Based on a comprehensive literature review and theoretical analysis, the research team preliminarily developed the fundamental framework for a scale measuring college students' knowledge and behavioral norms regarding the academic use of AI tools, adopting the Knowledge-Attitude-Behavior (KAB) model as its theoretical foundation. Using purposive sampling, the study recruited four students each from medical, science, and engineering disciplines at a selected university for semi-structured qualitative interviews. The inclusion criteria were: (1) experience in using generative AI tools for at least six months during their university studies, and (2) voluntary participation in the study. The interview questions included: (1) Based on your experience, what are your purposes and applications for using AI tools? (2) In your opinion, what behaviors associated with AI tool usage might violate academic integrity norms? After transcription and verification of the interview texts, the data were imported into Nvivo 14 for management and analyzed using a categorical analysis approach. Open coding was first employed to extract core categories related to AI tool usage purposes, behavioral patterns, and academic norms, thereby identifying the primary and secondary indicators for the scale. Subsequently, these categories were further refined to establish the tertiary indicators. Ultimately, the study developed a framework for the scale comprising two primary indicators, three secondary indicators, and eighteen tertiary indicators, forming a structured assessment of college students' knowledge and behavioral norms in the academic use of AI tools.

2.3. Development of the Consultation Questionnaire

Based on a literature review, semi-structured interviews, and group discussions, an expert consultation questionnaire was developed. This questionnaire consists of four main components: the questionnaire instructions, an expert demographic information form, an expert consultation form, and a survey assessing experts' familiarity with the research topic and their criteria for judgment. The expert consultation form adopts the Likert 5-point rating scale to evaluate the importance and operability of each indicator within the scale. Importance refers to the significance of an evaluation indicator in assessing university students' knowledge and behavior concerning the normative academic use of AI tools. The rating criteria are as follows: very important = 5 points, important = 4points, moderate=3 points, unimportant =2 points, and very unimportant =1 point. Operability refers to the feasibility of applying each indicator in practice, with the rating criteria as follows: excellent =5 points, good =4 points, moderate =3 points, poor = 2 points, and very poor=1 point. The expert demographic information form includes details such as gender, age, field of expertise, professional title, years of work experience, and highest educational qualification.

2.4. Implementation of the Delphi Expert Consultation

A purposive sampling method was employed to select experts for the Delphi consultation. Specifically, 11 experts from undergraduate institutions in Central China, South China, East China, and North China were chosen from various educational disciplines, including medicine, economics, computer science, and literature. Additionally, one expert from the University of South Wales in the United Kingdom, with a background in Chinese education, was included in the consultation panel. The consultation process involved two rounds conducted via email or WeChat. The inclusion criteria for experts were as follows: (1) a minimum of 10 years of experience in education or educational management, with a professional title of associate professor or higher; (2) possession of a master's degree or higher; and (3) informed consent and voluntary participation in the study. Through two rounds of expert consultation, the hierarchical indicators were further revised and refined.

2.5. Preliminary Questionnaire Survey

According to standard practice, the sample size should be at least 5 to 10 times the number of scale items [13]. In this study, a convenience sampling method was used to select Chinese 140 undergraduate nursing students as participants for the preliminary survey. Inclusion Criteria includes: Aged 18 years or older; Has used AI tools for ≥3 months; Willing to participate in this study; Possesses normal language comprehension skills and has no mental disorders. Exclusion Criteria includes: Younger than 18 years old; Has used AI tools for <3 months. The questionnaire was distributed and collected via Sojump (a Chinese online survey platform). Each IP address was restricted to a single submission, and only fully completed questionnaires were accepted for submission.

2.6. Formal Questionnaire Survey

Using a convenience sampling method and following inclusion and exclusion criteria, a questionnaire survey was conducted among 220 second- and third-year undergraduate nursing students at a university in Hunan, China, from September to October 2024. The researchers distributed the questionnaire links and instructions to the students via the Superstar Learning Platform. The survey instruments included a general demographic information questionnaire and a scale measuring university students' knowledge and behaviors regarding the academic use of AI tools. This study was reviewed and approved by the Ethics Committee of the University of South China (Approval No.: 20240077).

2.7. Statistical Methods

The questionnaire data were exported from Wenjuanxing and analyzed using SPSS 22.0 and AMOS 22.0. Descriptive statistics, including frequency, percentage, mean, and standard deviation, were used to summarize general information. The expert consultation results were analyzed using the coefficient of variation and Kendall's coefficient of concordance. The reliability and validity of the scale were assessed through item analysis, exploratory factor analysis, content consistency testing, and confirmatory factor analysis.

3. RESULTS

3.1. Expert Consultation Results

3.1.1. Expert Background and Authority

A total of 12 educational experts were selected to evaluate the scale in this study, including 3 males and 9 females. The experts' ages ranged from 38 to 60 years (mean ± standard deviation: 48.67 ± 7.98), and their years of work experience ranged from 10 to 40 years (mean ± standard deviation: 21.83 ± 10.34). Among them, 5 were associate professors, and 7 were full professors. Regarding their academic qualifications, 4 held master's degrees, and 8 held doctoral degrees. The response rates for the two rounds of expert consultation questionnaires were 91.67% (11/12) and 90.91% (10/11), respectively, indicating a high level of expert engagement. In the first round, the expert judgment consistency coefficient was 0.87, the familiarity coefficient was 0.84, and the expert authority coefficient was 0.85. In the second round, the expert judgment consistency coefficient was 0.88, the familiarity coefficient was 0.84, and the expert authority coefficient was 0.86.

3.1.2. Coordination of Expert Opinions

In the first round of expert consultation, the Kendall's coefficient of concordance for the overall importance rating was 0.542, and for the operability rating, it was 0.334. Both coefficients reached a significant level ($P = 0.000 < 0.001$). In the second round of expert consultation, the Kendall's coefficients for overall importance and operability increased to 0.598 and 0.560, respectively, both of which again reached a significant level ($P = 0.000 < 0.001$). The significant increase in Kendall's coefficients in the second round indicates an enhanced coordination of expert opinions, with greater consensus and reliable results.

3.1.3. Expert Consultation Results

The importance scores in the first round of expert consultation ranged from 3.09 to 5.00, with a coefficient of variation between 0.00 and 0.31. The operability scores ranged from 3.00 to 4.82, with a coefficient of variation between 0.08 and 0.37. In the second round of expert consultation, the importance scores ranged from 4.00 to 5.00, with a coefficient of variation between 0.00 and 0.16. The operability scores ranged from 4.00 to 5.00, with a coefficient of variation between 0.00 and 0.16. Consideration was given to removing items with an average importance score of ≤ 3.5 or a coefficient of variation > 0.25.

Based on expert opinions and statistical analysis results in the first round of consultation, and after group discussions, the following modifications were made: (1) Three tertiary indicators were deleted:

"*Using AI tools to correct grammar, polish, or provide structural suggestions for text creation constitutes academic misconduct*," "*I have used AI tools to modify the grammar, polish, or modify the structural framework of text,*" and "*I have used AI tools to substantially modify the text I created.*" (2) Some tertiary indicators were split, such as dividing "*I have used AI tools to complete exams*" into "*I have used AI tools to assist with open-book exams*" and "*I have used AI tools to assist with closed-book exams.*" (3) The language expression of certain tertiary indicators was simplified to make the meaning of each item clearer, for example, "*I know that using AI tools to ghostwrite coursework violates academic integrity principles*" was revised to "*Using AI tools to ghostwrite assignments constitutes academic misconduct*."

No changes were made to the indicators in the second round of expert consultation. After two rounds of expert consultations, the final framework of the scale was determined, which includes two primary indicators, three secondary indicators, and fifteen tertiary indicators. The "Usage Norms Knowledge" subscale corresponds to items H1 through H6, while the subscale "Academic Misconduct" encompasses academic tasks and academic achievements, which correspond to items F1 to F4 and F5 to F9, respectively (see Table 1).

Table 1. Initial Items of the University Students' Perception of Usefulness and Usage Behavior Scale for AI Tools

First-level Indicator	Items
Usage Norms Knowledge	H1. Using AI tools to ghostwrite assignments is considered academic misconduct.
	H2. Using AI tools to ghostwrite patents, papers, etc. is considered academic misconduct.
	H3. Using AI tools to significantly modify text without indicating the extent of assistance is considered academic misconduct.
	H4. Using AI tools to significantly modify text without indicating the extent of assistance is considered academic misconduct.
	H5. Using AI tools to modify or falsify research data is considered academic misconduct.
	H6. Using AI tools to assist in exams is considered academic misconduct.
Academic Misconduct	F1. I have used AI tools to ghostwrite assignments.
	F2. I have used AI tools to significantly modify assignments.
	F3. I have cited content generated by AI without indicating the source.
	F4. I have cited references and data generated by AI in my writing.
	F5. I have used AI tools to ghostwrite academic projects and achievements such as patents and papers.
	F6. I have used AI tools to significantly modify academic projects and achievements such as patents and papers.
	F7. I have used AI tools to adjust data to obtain better research results.
	F8. I have used AI tools to assist in open-book exams.
	F9. I have used AI tools to assist in closed-book exams.

3.2. Pre-test and Initial Scale Refinement Results

3.2.1. Pre-test

A total of 140 undergraduate nursing students were surveyed in the pre-test. The researchers excluded questionnaires with excessively long or short response times, as well as those with answers showing clear patterns or logical inconsistencies. Twelve questionnaires were discarded, leaving 128 valid responses, resulting in a valid response rate of 91.43%. The general demographic information of the pre-test participants is shown in Table 2. The purposes of AI tool usage are diverse, with only the five most common purposes for nursing students being listed.

Table 2. General Information of Pre-test and Formal Test Subjects

Item	Category	Pre-test	Formal Test
Gender	Male	13 (10.2)	21 (10.5)
	Female	115 (89.9)	179 (89.5)
Grade	Sophomore	47 (36.7)	88 (44.0)
	Junior	81 (63.3)	112 (56.0)
Duration of Use	≤6 months	48 (37.5)	93 (46.5)
	7~12 months	44 (34.3)	61 (30.5)
	13~18 months	14 (11.0)	16 (8.0)
	19~24 months	24 (11.0)	19 (9.5)
	25~30 months	3 (2.4)	4 (2.0)
Purpose of Use	≥31 months	4 (3.1)	7 (3.5)
	Improve course learning efficiency	105 (82.0)	85 (14.9)
	Knowledge understanding and integration	42 (19.4)	95 (16.6)
	Find innovative inspiration or ideas	85 (66.4)	75 (13.1)
	Improve writing efficiency	80 (62.5)	88 (15.4)
	Improve writing quality	64 (50.0)	65 (11.4)
	Assist in completing assignments	65 (50.8)	61 (10.7)
	Expand knowledge outside the major	50 (39.1)	57 (10.0)
	Learn AI technology	34 (26.6)	27 (4.7)

3.2.2. Initial Scale Refinement Results

Table 3. Exploratory Factor Analysis Results

Common Factor	Item	Commonality Coefficient	Rotated Factor Loadings		
			1	2	3
Usage Norms Knowledge	H1	0.599	0.774	0.012	0.016
	H2	0.788	0.883	0.036	0.085
	H3	0.709	0.838	0.003	0.079
	H4	0.693	0.824	0.089	0.072
	H5	0.810	0.900	0.012	0.022
	H6	0.821	0.906	0.003	0.018
Academic Tasks Misconduct	F1	0.829	0.138	0.164	0.885
	F2	0.711	0.118	0.232	0.802
	F3	0.554	0.122	0.443	0.586
Academic Achievements Misconduct	F5	0.798	0.024	0.836	0.314
	F6	0.829	0.013	0.895	0.164
	F7	0.410	0.102	0.609	0.168
	F8	0.555	0.069	0.730	0.132
	F9	0.818	0.074	0.901	0.038
Eigenvalues			4.529	4.156	1.238
Cumulative Variance Contribution Rate			31.858	56.871	70.878

First, the extreme value method was used to compare the differences between high and low groups on each item, and independent sample T-tests were conducted to test whether the differences between high and low groups on each item in the scale were statistically significant. If the absolute value of T was <3.0 or P≥0.05, the item was deleted [13]. After analysis, F4 had a T value of 1.849 and P=0.069, so it was deleted. Second, the remaining items were analyzed for correlation, and the correlation coefficients between each item and the overall scale score and the corresponding subscale scores ranged from 0.394 to 0.654, 0.792 to 0.898, and 0.640 to 0.851, respectively, with P values <0.001, indicating significant correlations. All 14 items were retained. Second, Cronbach's α coefficient and corrected item-total correlation (CITC) were used to analyze the overall reliability and item reliability of the initial scale. The Cronbach's α coefficients for the total scale and subscales were 0.805, 0.925, and 0.855, respectively, indicating high overall reliability and acceptability. The CITC values for the total scale items and subscale items ranged from 0.312 to 0.580, 0.685 to 0.851, and 0.495 to 0.793, respectively, indicating that the quality of each item was acceptable [13], and no items needed to be deleted. Then, KMO and Bartlett's sphericity tests were conducted to assess the suitability of the data for factor analysis. The results showed that the KMO value for the total scale was 0.827, and Bartlett's sphericity test χ^2 was 1191.063, df=91, significance level $P=0.000<0.05$, indicating that the data were suitable for factor analysis. Finally, principal component analysis (PCA) was used to perform factor analysis on the total scale items, extracting factors based on eigenvalues greater than 1, and using the maximum variance method (Varimax) for orthogonal rotation. According to the standards [13], if the commonality coefficient was ≤0.2 or the factor loading was <0.45, the item was deleted. The results showed that the commonality coefficients ranged from 0.410 to 0.829, and the factor loadings ranged from 0.659 to 0.821, with a cumulative variance contribution rate of 70.878% (see Table 3), which is higher than the acceptable standard of 60%. Based on the above analysis, a scale with 14 items was finally formed.

3.3. Formal Test Results

3.3.1. General Information of Formal Test Subjects

A total of 220 questionnaires were collected in the formal test, and 20 invalid questionnaires were excluded, resulting in 200 valid questionnaires, with a valid response rate of 90.09%. The general information of the test subjects is shown in Table 2.

3.3.2. Confirmatory Factor Analysis (CFA)

AMOS was used for confirmatory factor analysis to further test the stability of the scale structure. The results (see Table 4) showed: $\chi 2/df$ was 2.978, GFI was 0.859, PGFI was 0.606, RMSEA was 0.098, RMR was 0.048, IFI was 0.921, TLI was 0.902, CFI was 0.920, PCFI was 0.748, PNFI was 0.721, and the final three-dimensional structure with 14 items was obtained, as shown in Figure 1.

Table 4. Confirmatory Factor Analysis Results

Common Factor	Item	Unstandardized Factor Loading	Standardized Factor Loading	CR Value
Usage Norms Knowledge	H1	1.000	0.741	
(Common Factor 1)	H2	1.030	0.862	12.600***
	H3	0.962	0.811	11.772***
	H4	0.971	0.777	11.222***
	H5	1.089	0.896	13.152***
	H6	1.073	0.906	13.318***
Academic Tasks Misconduct	F1	1.000	0.756	
(Common Factor 2)	F2	0.813	0.639	8.046***
	F3	0.957	0.817	9.184***
Academic Achievements Misconduct	F5	1.000	0.909	
	F6	0.759	0.633	10.373***
(Common Factor 3)	F7	0.815	0.834	16.539***
	F8	0.910	0.927	20.589***
	F9	0.695	0.578	9.153***
Note: *** indicates P < 0.001				

3.3.3. Competitive Model Analysis

A competitive model strategy was used to test whether the three-dimensional structure of the scale on college students' perception of AI tool usefulness and usage behavior was the optimal model. The results showed that the fit indices of the three-factor model were overall better than those of the single-factor and two-factor models, demonstrating higher model fit and explanatory power, as shown in Table 5.

Table 5. Scale on College Students' Perception of AI Tool Usefulness and Usage Behavior

Model	$\chi 2$/df	GFI	RMSEA	CFI	IFI	TLI
Single-Factor Model	13.729	0.483	0.253	0.484	0.487	0.390
Two-Factor Model	4.305	0.803	0.129	0.868	0.869	0.842
Three-Factor Model	2.978	0.859	0.098	0.920	0.921	0.902

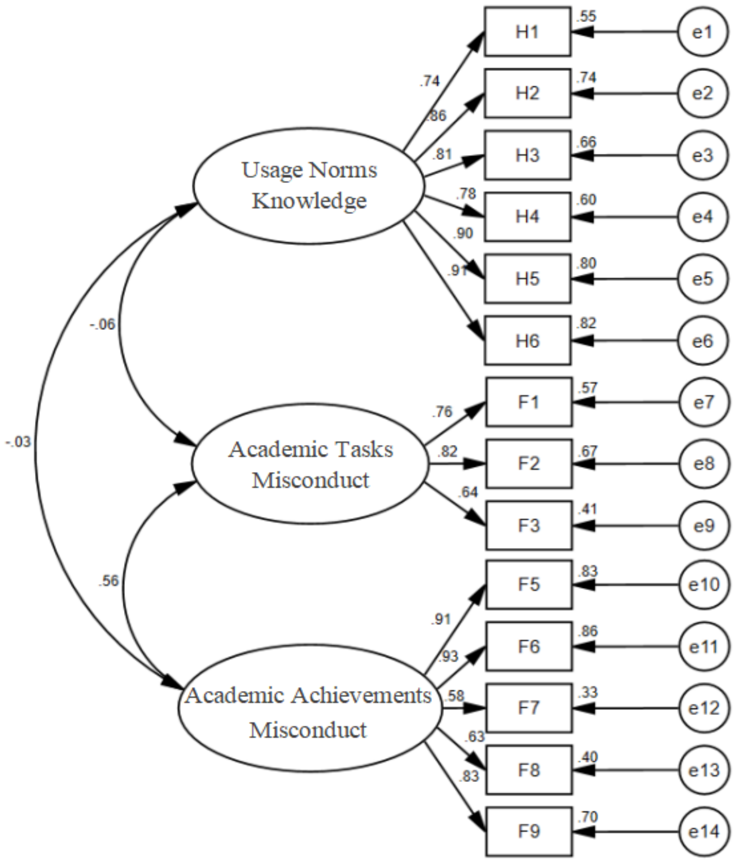

Figure 1. Confirmatory Factor Analysis Results

3.3.4. Reliability Test

The formal test analysis results showed that the Cronbach's α coefficients for the scale and subscales were 0.811, 0.930, and 0.865, respectively; the Cronbach's α coefficient for the usage norms dimension was 0.930, and the composite reliability (CR) value was 0.932; the Cronbach's α coefficient for the academic tasks dimension was 0.775, and the CR value was 0.783; the Cronbach's α coefficient for the academic achievements dimension was 0.873, and the CR value was 0.889, all higher than the acceptable reliability standards. In addition, considering the heterogeneity between the "Knowledge" and "Academic Misconduct Behavior" subscales, this heterogeneity would weaken the internal consistency of the total scale, thereby reducing the split-half reliability of the total scale. Therefore, this study only reports the split-half reliability coefficients of the 2 subscales, which were 0.891 and 0.860, respectively, indicating good split-half reliability of the scale.

3.3.5. Structural Validity Test

The analysis results showed that the standardized loadings of the items in the three factors of the scale ranged from 0.741 to 0.906, 0.639 to 0.817, and 0.578 to 0.927, respectively, all exceeding the minimum standard of 0.5 and reaching a significant level; the AVE values of the three factors were 0.696, 0.549, and 0.623, respectively, all > 0.5, indicating good convergent validity of the scale. In addition, the square roots of all factor AVEs were greater than their corresponding correlation coefficients, indicating good discriminant validity of the scale.

4. DISCUSSION

4.1. The Rationality and Scientific Validity of the Scale for Academic Use of AI Tools by University Students

This study, based on the KAB theory, employed a combination of methods including literature review, semi-structured interviews, Delphi expert consultation, and questionnaire surveys, integrating both qualitative and quantitative research results, to develop a scale for assessing university students' knowledge and behaviors regarding the academic use of AI tools. Through literature review and semi-structured interviews, expert consultation indicators at three levels were determined. The results of expert consultation in this study were highly authoritative and reliable, and the questionnaire demonstrated good content validity, leading to further revisions and optimization of the scale items. This ensured the scientific and rational nature of the scale content: (1) The effective recovery rates of the expert consultation questionnaires in the first and second rounds were 91.67% and 90.91%, respectively, indicating a high level of expert engagement in the study. The expert authority coefficients for the two rounds were 0.85 and 0.86, respectively. (2) The Kendall concordance coefficients for importance and reliability in the first round of expert consultation were 0.542 and 0.334, respectively, while in the second round, these values increased to 0.598 and 0.560, with both reaching a significance level of $P < 0.001$, indicating a consensus among expert opinions.

4.2. The Academic Use Norms and Behavioral Scale for AI Tools among College Students Demonstrates Reliability

This study evaluated the reliability and validity of the scale through pilot testing and formal testing. Item analysis results indicate that each item demonstrates good discrimination and homogeneity. Exploratory factor analysis revealed that the scale comprises three common factors with eigenvalues greater than 1, accounting for a cumulative variance contribution of 70.878%, with all items loading onto the predefined dimensions. Confirmatory factor analysis confirmed that the scale exhibits a good model fit, suggesting strong structural validity. Internal consistency analysis revealed that the Cronbach's α coefficients for the total scale and subscales were 0.811, 0.930, and 0.865, respectively, while the split-half reliability coefficients for the subscales were 0.891 and 0.860, indicating high internal consistency. Therefore, this scale is reliable for measuring college students' academic use norms and behaviors related to AI tools and provides a dependable measurement tool for related research.

4.3. The academic use norms and behavioral scale for AI tools among university students holds practical significance

The data from this study indicates that an increasing number of nursing students are utilizing generative AI tools for learning, knowledge comprehension and expansion, completing various writing tasks, and seeking innovative ideas and insights. The integration of generative AI technologies and their tools into the educational system has become an inevitable trend and will serve as a significant force in driving higher education reforms. However, despite the potential value that AI tools can offer in research ideation, excessive reliance on generative AI tools, coupled with a tendency to directly copy and paste generated content, may diminish students' opportunities for effective practice in creative and critical thinking skills, thereby hindering the development of creativity and critical thinking. Consequently, educators must fully recognize the potential risks associated with the use of AI tools, promptly monitor possible academic misconduct arising from students' use of these tools, and establish appropriate usage norms to constrain inappropriate behavior, thereby promoting the effective educational empowerment of AI tools. The development of this scale aims to assist educators in assessing students' awareness and behavioral performance regarding academic use norms for AI tools. With results-based evaluations, timely and effective measures can be implemented to prevent academic misconduct and guide students in leveraging AI tools to enhance personalized learning while fostering their critical thinking and innovative capabilities.

5. CONCLUSION

The academic use guidelines and behavior scale for AI tools developed in this study consists of 14 items, covering three dimensions: usage norms, academic tasks, and academic performance. In the reliability and validity analysis, the scale demonstrated good internal consistency and structural validity, exhibiting high reliability and scientific rigor. Therefore, educators can use this scale to assess university students' knowledge of academic usage norms for AI tools and potential academic misconduct, providing a basis for educational interventions.

6. LIMITATIONS AND FUTURE RESEARCH

However, the study's sample primarily consisted of nursing students from a specific university and did not include senior students who are required to write graduation theses, which may impose certain limitations on the data reporting for individual items under the "academic performance" dimension. Additionally, this study lacks a comparative analysis between students from different disciplines and institutions.

Future research will expand to include students from various disciplines and educational levels, further enriching the scale's items, enhancing its robustness and applicability, and laying the foundation for the development of academic use guidelines for AI tools. This will promote the effective empowerment of AI technology in education, support the personalized development of university students, and foster the development of critical thinking, innovation, and academic integrity.

CONFLICTS OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ACKNOWLEDGEMENTS

We are extremely grateful to the participants who participated in this study and made this research possible. We would also like to thank the Chuanshan College of University of South China for their grant support (Grant No. 2024CY001).

REFERENCES

[1] Tertiary Education Quality and Standards Agency, (2022). What is academic integrity? https://www.teqsa.gov.au/students/understanding-academic-integrity/what-academic-integrity

[2] Jereb, E., Perc, M., Lämmlein, B., Jerebic, J., Urh, M., Podbregar, I., & Šprajc, P. (2018). Factors influencing plagiarism in higher education: A comparison of German and Slovene students. PLOS ONE, 13(8), e0202252. https://doi.org/10.1371/journal.pone.0202252

[3] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine, 183(6), 589. https://doi.org/10.1001/jamainternmed.2023.1838

[4] Liu, H., Azam, M., Bin Naeem, S., & Faiola, A. (2023). An overview of the capabilities of ChatGPT for medical writing and its implications for academic integrity. Health Information & Libraries Journal, 40(4), 440–446. https://doi.org/10.1111/hir.12509

[5] Chris Westfall (2023). Educators Battle Plagiarism As 89% Of Students Admit To Using OpenAI's ChatGPT For Homework [N/OL]. Forbes. https://www.forbes.com/sites/chriswestfall/2023/01/28/educators-battle-plagiarism-as-89-of-students-admit-to-using-open-ais-chatgpt-for-homework/?sh=12e01a42750d

[6] Sun, G. H., & Hoelscher, S. H. (2023). The ChatGPT Storm and What Faculty Can Do. Nurse Educator, 48(3), 119–124. https://doi.org/10.1097/NNE.0000000000001390

[7] Ministry of Foreign Affairs (2022). China's Position Paper on Strengthening Ethical Governance of Artificial Intelligence. https://www.mfa.gov.cn/web/ziliao_674904/zcwj_674915/202211/t20221117_10976728.shtml

[8] Yang, J.F., Chu, J., Zhang B.X.(2024). Research on Constructing Ethical Norm Indicators for the Application of Artificial Intelligence in Education [J]. E-education Research. DOI:10.13811/j.cnki.eer.2024.10.003

[9] Hong, W. C. H., Chi, C., Liu, J., Zhang, Y., Lei, V. N.-L., & Xu, X. (2023). The influence of social education level on cybersecurity awareness and behaviour: A comparative study of university students and working graduates. Education and Information Technologies, 28(1), 439–470. https://doi.org/10.1007/s10639-022-11121-5

[10] Fresán, A., Domínguez-Martínez, T., Castilla-Peón, M. F., Robles, R., Hernández, O., Vélez, T., & Muñoz, C. (2022). Confirmatory Factor Analysis of the Transgender Knowledge, Attitudes, and Beliefs (T-KAB) Scale for the Mexican Population. Archives of Sexual Behavior, 51(4), 1959–1966. https://doi.org/10.1007/s10508-021-02259-3

[11] Chory, A., Nyandiko, W., Beigon, W., Aluoch, J., Ashimosi, C., Munyoro, D., Scanlon, M., Apondi, E., & Vreeman, R. (2021). Perspectives of education sector stakeholders on a teacher training module to reduce HIV/AIDS stigma in Western Kenya. BMC Public Health, 21(1), 1281. https://doi.org/10.1186/s12889-021-11331-5

[12] Chen Xiangming (1999). Classification and Analysis of Data [J]. Social Science Front, 223-229.

[13] Wu, M.L. (2010). Practical Statistical Analysis of Questionnaires: SPSS Operations and Applications [M]. Chongqing: Chongqing University Press.

[14] Nachalon, Y., Broer, M., & Nativ-Zeltzer, N. (2024). Using ChatGPT to Generate Research Ideas in Dysphagia: A Pilot Study. Dysphagia, 39(3), 407–411. https://doi.org/10.1007/s00455-023-10623-9

[15] Nanzhong, W., Xianzhang, C., Yong, F.(2023). From "Disordered" to "Ordered": The Transformation and Its Generation Mechanism of Generative Artificial Intelligence Education Application.Journal of Distance Education, 6:42-51. https://doi.org/10.15881/j.cnki.cn33-1304/g4.2023.06.005