Canadian Journal of Educational and Social Studies
Vol. 6(2), 2026, pp. 59-71

Human vs. AI: Assessing Scale Development for Perceived Risks of ChatGPT in Academic Settings

Bouzar Abdelouahd[1] & Khaoula EL Idrissi[2] & Samia Moustaghfir[3]

[1,2] Department of English Studies, Sidi Mohammed Ben Abdellah University, Fez, Morocco
[3] Samia Moustaghfir, Arts and Pedagogical Engineering Research Laboratory, Faculty of Languages, Letters and Arts, Ibn Tofail University, Kenitra, Morocco
Correspondence: Bouzar Abdelouahd, Sidi Mohammed Ben Abdellah University, Fez, Morocco
Email: abdelouahd.bouzar@usmba.ac.ma; khaoula.elidrissi@usmba.ac.ma; samia.moustaghfir@uit.ac.ma

DOI: 10.53103/cjess.v6i2.469

Abstract

In this exploration, we conducted an in-depth comparative research of psychometric instrument development and compared the potential of human subject-matter professionals with artificial intelligence systems to conduct the analysis of perceived risks related to ChatGPT in the academic setting. The respondents (two professor-level experts on research methodology) and the AI systems ChatGPT and Claude, produced four different eighteen-item scales, all of them dealing with the triad of risk measures: psychological, ethical, and practical. A group of twenty experienced evaluators was used to evaluate the quality of the items using the accepted assessments of evaluation after which the resulting scales were provided to a sample of 120 respondents who represented a wide academic speciality. Further statistical analyses involved Exploratory Factor Analysis (EFA), as well as, measurement of reliability through Cronbach alpha, calculation of Average Variance Extracted (AVE) and calculation of composite reliability. Evidence showed that the scales produced by AI had significantly higher psychometric quality; in particular, the scale of Claude had the highest score of reliability with the $\alpha = 0.942$ and the composite reliability value CR= 0.930. The ANOVA did not show statistically significant differences between the scales ($F < 2.183$, $p < 0.168$), however, the effect size analysis placed significant emphasis on some differences in the scales; in particular, the effect size obtained between Claude and Professor 1 was large ($d < -1.578$). Measurement Factor loading tests supported the construct validity of all measures, but AI-generated items indicated a slightly better factor structure. This further implied that the AI-generated scales, and particularly the scale created by Claude, were better than the human-created scales in regards to Clarity, Specificity and Overall Quality, but the human-created scales retained competition in regards to Relevance. All these outcomes mean that AI systems can produce high-quality psychometric tools that are equal or even superior to conventional human-created scales, meaning that they could change the effectiveness of instrument creation in psychological research and remain at high psychometric standards.

Keywords: Psychometric Scales, Reliability Assessment, Factor Analysis, Applications of Artificial

Intelligence, Educational Technology

Introduction

Scale development is an essential element of empirical research in the social sciences and psychology, which enables the quantification of latent constructs that can not be directly observed (Raykov and Marcoulides, 2011). Such careful development of psychometric scales that are reliable and valid requires deep knowledge on the conceptualization of constructs, creation of items, extensive statistical testing, and validation of the scales and all this is based on the known principles of psychometrics theory (Nunnally & Bernstein, 1994). In the past, this task has relied on experienced scholars who employ theoretical acumen, linguistic acuity and methodological strictness in creating items that achieve construct validity, clarity, specificity, and avoid the trap of the double-barreled statement (Hinkin, 1998; Worthington and Whittaker, 2006). These professionals also need to predict response patterns, cultural peculiarities, and possible bias when using powerful statistical methods, such as structural equation modeling, to checking the scale structures (Kline, 2016).

The abrupt introduction of artificial intelligence applications and especially big language models like ChatGPT and Claude into new academic environments is what has spawned the apprehension about the integrity of scholarly practices, teaching efficacy, and the overall maturation of students (Cotton et al., 2023). The assumed risks include psychological aspects (such as a decline in critical thinking due to the overuse of AI), ethical issues (plagiarism, dishonesty in academic aspects), and practical challenges (validity and fair accessibility of AI-created content). These multifaceted risks are to be assessed with advanced psychometric tools that are specifically designed to be used in scholastic use of ChatGPT, but the available scales are mostly about generic technology acceptance, thus a visible gap in terms of specific risk assessment assessments would be present.

Recent advances in natural language processing have shed light on the ability of artificial intelligence to perform complex tasks previously considered as belonging to human knowledge (Brown et al., 2020). Large language models also exhibit the ability to create contextually suitable material and solve complex instructions, raising the question of their application to the psychometric scale development area, which requires a more delicate understanding of the theory of measurement and the properties of items (Nunnally and Bernstein, 1994). Though it has this developing potential, there are few systematic comparative analyses of the psychometric integrity of AI generated scales over human developed scales, especially in relation to constructs associated with AI use.

The current research attempts to fill this gap by comparing scales developed by two professors of research methodology, ChatGPT, and Claude, to assess purported risks of ChatGPT in academic contexts on psychological, ethical, and practical levels. The quality of scale is evaluated using already defined psychometric values, including reliability (Cronbachalpha; Tavakol and Dennick, 2011), signs of validity (Average Variance Extracted), and the sufficiency of factor structure, using the methods of Exploratory Factor Analysis, and structural equation modelling (Revelle, 2021; Rosseel, 2012). Supplementary professional ratings also judge clarity, relevance and specificity of items and follow best practice in scale development (Worthington and Whittaker, 2006). The study hopes to find out whether AI systems can create psychometric tools that will match, or even exceed human-designed scales and this may lead to a revolution in the efficiency of scale development, on the other hand, maintaining high standards of rigor psychometrically.

Research Objective

The main aim of the given study is to compare the psychometric quality of the scales created by two professors, ChatGPT, and Claude to assess perceived risks (psychological, ethical, practical) related to the use of ChatGPT in the academic setting. Their reliability and validity are assessed with the help of extensive psychometric measures such as Cronbachs alpha, Average Variance Extracted (AVE), composite reliability (rho A) and factor loadings, with the aim being to determine whether AI generated scales can be as good or better than those created by human researchers.

Methods

Scale Development

The scale creation process included four autonomous developers, two specialists in research methodology at two different universities and two artificial intelligence systems (ChatGPT and Claude). All developers were requested to come up with an 18-item instrument to measure perceived risks related to using ChatGPT as an academic resource, six items were assigned to the following areas: psychological risks (e.g., over-reliance, loss of critical thinking), ethical risks (ex: academic integrity, plagiarism), and practical risks (e.g., accuracy, availability). Each of the items was designed to be answered in seven points Likert scale, with 1 (Strongly Disagree) and 7 (Strongly Agree).

The AI systems were carefully designed to ensure uniformity in the instructions: the scale was designed with six items per group (e.g. psychological (e.g. over-reliance), ethical (e.g. academic integrity, plagiarism), and practical (e.g. accuracy, accessibility)) based on clearly-written and short statements that are fit within a seven-point Likert scale. Human developers were given the same written instructions, including same construct definitions and item requirements.

To be able to produce high quality and relevant items, pilot testing of the AI prompts was implemented before finalizing the scales. Several iterations of each AI system were created and the best iteration was chosen by preliminary evaluation in terms of grammatical correctness, readability, and correspondence to the target constructs. Each of the items was first reviewed to ensure that it was grammatically correct, understandable and construct valid before undergoing an expert review.

Expert Evaluation

A total of twenty professional scholars in research methodology were recruited to rate all items of the scale anonymously, and represent a diversity of academic departments. The professionals used a 7-point Likert scale to evaluate every item over five dimensions, i.e., clarity, relevance, specificity, the absence of the double-barrel statements, and quality. An elaborate rubric was used to establish clear scoring guidelines on every dimension thus maintaining consistency in ratings.

In a way to reduce evaluator fatigue, the 72 items (which represent 18 items in four scales) were divided into several evaluation sessions. All 20 professors rated each item, thus ensuring that there was enough data to support reliability testing. Inter-rater reliability was measured using the Intraclass Correlation Coefficient (ICC) measure to assess the consistency of evaluators.

Products with a mean score less than 4 out of 7 or less items in the first quartile were labelled as may be dropped. However, at least three to four items per construct per scale were kept in order to have reasonable construct coverage in further analysis.

Pilot Study

A pilot study was conducted with 25 individuals in order to investigate the clarity of items, the understanding of the respondents, and the primary psychometric characteristics. Feedback received during pilot participants had provided clues to minor adjustments to the wording of the items and the actual administration of the survey before the actual study.

Main Study

The polished instruments were given to 120 respondents who were selected in various academic departments in different universities. The stratified sampling ensured even representation of disciplines that included STEM fields, social sciences, humanities, business, health sciences and education. The questionnaire was shared on the Qualtrics platform, and recruitment was supported by Prolific, which increased the sample diversity and increased the generalizability. Procedures related to ethical compliance involved obtaining informed consent, ensuring anonymity of the respondents, and following up on the institutional review board procedures. The respondents provided demographic data, such as academic field, level of study, and experience using ChatGPT before to allow the stratified analyses.

Analysis

The EFA was implemented using the principal axis factoring with oblimin rotation with the assumption of correlated factors that portray three risk dimensions. Factors that have a factor loading of less than 0.4 were thought of as being discarded, and a balance between interpretability and the retention of the item was achieved according to the set psychometric standards.

The measures of reliability included the calculation of Cronbach alpha of each subscale where the value of 0.7 was accepted as the level of internal consistency. Average Variance Extracted (AVE) was calculated and a cut off of 0.5 taken, convergent validity was obtained by making sure that AVE was greater than squared inter-factor correlations. Additional reliability tests were given by composite reliability and rho a coefficients.

The psychometric metrics of the statistical comparisons between scales were done using Analysis of Variance (ANOVA) and t-tests with the effect sizes (Cohen d) reporting practical significance. Practical differences were also evaluated as per analyzing item interpretability and responding to feedbacks of respondents. Items of AI and human items were confronted in terms of tone, complexity, and specificity to identify possible systematic variation.

Qualtrics was used to administer the surveys on survey and Prolific to recruit the participants. This was analyzed statistically using the R programming language under the following specific packages: the irr package to calculate intraclass correlation coefficient, the psych package to analyze EFA and reliability, the semTools package to calculate AVE and composite reliability, the lavaan package to perform confirmatory factor analysis, and the ggplot2 package to visualize data.

Results

Participant Demographics

The last sample consisted of 120 participants who had different academic backgrounds. The sample consisted of undergraduate students (40 per cent), graduate students (35 per cent), and doctoral students (25 per cent) in six broad disciplinary areas: STEM (25 per cent), Social Sciences and Humanities (38 per cent), Economics (37 per cent). The participants were between 18 and 45 years (M 26.3, SD 6.8), and their ChatGPT experience was Limited (35%), Moderate (50%), and Extensive (15%).

Table 1: Sample demographic profile

Category	Subcategory	Count	Percentage
Total Sample	All Participants	120	100%
Academic Level	Undergraduate Students	48	40%
	Graduate Students	42	35%
	Doctoral Students	30	25%
Disciplinary Areas	STEM	30	25%
	Social Sciences & Humanities	46	38%
	Economics	44	37%
Age Demographics	Range	18-45 years	-
	Mean (M)	26.3 years	-
	Standard Deviation (SD)	6.8 years	-
ChatGPT Experience	Limited	42	35%
	Moderate	60	50%
	Extensive	18	15%

Expert Evaluation Results

Inter-rater reliability evaluation showed that there were moderate to high levels of agreement between all evaluators of the experts in most criteria. ICC ranged between 0.112 and 0.317, with specificity recording the highest agreement (ICC= 0.317), and avoidance of double-barreled statements registering the lowest (ICC= 0.112).

Table 2: Inter-Rater reliability (ICC) for evaluation criteria

Evaluation Criterion	ICC Value	95% CI	Interpretation
Clarity	0.135	[0.089, 0.196]	Fair
Relevance	0.140	[0.093, 0.202]	Fair
Specificity	0.317	[0.245, 0.401]	Moderate
Avoidance of Double-Barreled	0.112	[0.071, 0.167]	Fair
Overall Quality	0.117	[0.075, 0.173]	Fair

According to the experts, 63 out of 72 initial items were left to be used in the further analysis. Scale-wise, there was a balance in the number of items retained; Scale A (Professor 1) retained 15 items, Scale 2 (Professor 2) retained 16 items, Scale 3 (ChatGPT) retained 16 items, and Scale 4 (Claude) retained 16 items. Each of the scales was sufficiently represented in the three risk constructs.

Table 3: Comparison of scales across expert evaluation criteria

Scale	Developer	Clarity	Relevance	Specificity	Avoidance of Double-Barreled	Overall Quality
A	Professor 1	5.8	6.2	5.6	5.7	5.8
B	Professor 2	6.0	5.8	5.9	6.2	5.9
C	ChatGPT	6.4	6.3	6.6	5.6	6.3
D	Claude	6.5	6.4	6.5	6.3	6.5

Moreover, in Table 3, AI-generated scales tended to be higher in most criteria compared to those created by human beings, where Claude (D) and ChatGPT (C) were top in Specificity and Generality, respectively. Professor 2 (B) was higher in Avoidance and Relevance, whereas, Professor 1 (A) scored lower on criteria. These findings, as corroborated by the higher reliability (Table 4) and factor loading (Table 7) of AI scales, indicate that they are more precise and high in quality although human scales are competitive in the Relevance, reflecting balanced item retention, and the potential of AI in scale development.

Factor Analysis Results

The existence of a three-factor configuration, defined as psychological, ethical, and practical risk dimensions, was supported in all of the instruments assessed by exploratory factor analysis. Kaiser-Meyer-Olkin measures have satisfied the requirement of good sampling adequacy with a value of 0.918 to 0.947 being known to be good. The test of sphericity by Bartlett showed significant values in all the scales ($p < 0.001$) thus supporting the appropriateness of carrying out a factor analytic approach.

Table 4: KMO, Bartlett’s test, and variance explained across developers

Scale	Developer	KMO	Bartlett's p-value	Variance Explained (%)
A	Professor 1	0.918	< 0.001	68.2
B	Professor 2	0.927	< 0.001	71.4
C	ChatGPT	0.942	< 0.001	69.8
D	Claude	0.947	< 0.001	72.1

Reliability and Validity Analysis

Instruments all had excellent psychometric qualities as they were reflected in a variety of indices of reliability and validity. The alpha coefficients of Cronbach exceeded the 0.9 mark in all the scales, which indicated excellent internal consistency. The scale provided by Claude had the best total reliability (0.942), which was followed by ChatGPT

(0.936), Professor 2 (0.927) and Professor 1 (0.913).

Table 5: Reliability and validity indicators for study measurement instruments

Scale	Cronbach's α	Average AVE	Composite Reliability	Kaiser-Meyer-Olkin (KMO)
A (Professor 1)	0.913	0.657	0.894	0.918
B (Professor 2)	0.927	0.691	0.904	0.927
C (ChatGPT)	0.936	0.672	0.925	0.924
D (Claude)	0.942	0.689	0.930	0.947

Construct-level analysis indicated that there were consistent patterns on the risk dimensions. All of the constructs achieved high reliability ($\geq$ 0.89) and the items that proved to be the most reliable were ethical risk items (mean 0.936-0.942). However, psychological and practical risk constructs were found to be more varied across scales with AI-generated scales typically producing higher degrees of reliability coefficients.

Table 6: Cronbach's Alpha by construct and developer

Developer	Ethical	Practical	Psychological
Professor 1	0.940	0.891	0.909
Professor 2	0.936	0.915	0.925
ChatGPT	0.941	0.932	0.936
Claude	0.942	0.936	0.925

Statistical Comparisons

The comparison of Cronbach's alpha values of the various developers (Claude, ChatGPT, Professor 1 and Professor 2) showed that there were no statistically significant differences in reliability between them (F = 2.183, p = 0.168) so that the apparent differences may be due to the random variation between the developers. However, pairwise effect-size tests that used Cohen d showed significant practical differences between some pairs of developers. An example of such comparisons includes Claude versus Professor 1 (d = -1.578) and ChatGPT versus Professor 1 (d = -1.245) where the results obtained were large in effect size, which suggests that non-significant results are more meaningful in practice. Other comparisons had medium (Professor 2 vs Professor 1, d = -0.754) and small (Claude vs ChatGPT, d = -0.333) effects. These results highlight the need to consider statistical and practical significance in judging scale reliability in different developers, as per traditional scales (|human|>These findings clarify the significance of judging scale reliability based on both statistical and practical significance in industries with different developers. In turn, the findings indicate that the scales developed by AI can have higher internal reliability than those created by individual professors and, thus, should be investigated further in the upcoming research.

Table 7: Mean differences and effect sizes

Comparison	Mean Difference	Cohen's d	Effect Size Interpretation
Claude vs. Professor 1	0.029	-1.578	Large
ChatGPT vs. Professor 1	0.023	-1.245	Large
Claude vs. Professor 2	0.021	-1.456	Large
ChatGPT vs. Professor 2	0.019	-1.387	Large
Professor 2 vs. Professor 1	0.014	-0.754	Medium
Claude vs. ChatGPT	0.006	-0.333	Small

Factor Loading Analysis

Factor-loading analysis showed that itemconstruct relations were very strong in all scales. The percentage of items with increments of 0.7 or more differed between scales, with ChatGPT having the highest percentage (100 percent of items), Claude (94 percent), Professor 1 (93 percent), and Professor 2 (75 percent). The mean maximal factor loadings were between 0.811 and 0.830, which means that the construct validity of all scales is high.

Table 8: Summary of factor loadings by scale and developer

Scale	Developer	Items ≥ 0.7 Loading	Percentage	Average Max Loading
A	Professor 1	14/15	93%	0.811
B	Professor 2	12/16	75%	0.830
C	ChatGPT	16/16	100%	0.818
D	Claude	15/16	94%	0.828

Discussion

Scale Performance Comparison

The findings indicate that AI-generated scales, especially Claude scale, had high psychometric properties compared to human-created scales in various assessment measures. Claude scale had the best reliability (alpha = 0.942), composite reliability (CR = 0.930), and sampling adequacy (KMO = 0.947) in addition to achieving a very good factor structure with 94 percent of items showing a high factor loading (= 0.7). The size of ChatGPT worked similarly, which implied that modern AI systems have advanced psychometric scale development systems.

These results contradict traditional beliefs about the inability of scale development without the role of human expertise. Although human developers are providing good theoretical knowledge and intuitive idea of the principle of measurement, AI systems have been shown to perform similarly or better when it comes to producing items with high psychometric qualities. The stability of AI when applied to various constructs (psychological, ethical, and practical risks) suggests that researchers have strong understanding of measurement principles and not a local achievement in certain areas.

Implications for Psychometric Research

The fact that AI-generated scales perform better has far-reaching consequences to psychometrics research and practice. Potential benefits of AI systems include uniformity in the items generated, absence of human cognitive biases, the ability to generate large quantities of alternative items in a relatively short period, and future cost and time reductions in scale-development projects. Such capabilities have the potential to make scale development democratic in the sense that they make high-quality instrument development accessible to researchers who do not necessarily have a high level of psychometric training.

Nevertheless, all the results also raise serious questions of the quality of expertise in scale development. The conventional focus on theoretical knowledge and methodological training could be in need of an appreciation of AI possibilities and limitations. Hybrid solutions that would integrate the efficiency of AI and involve human control and conceptual basis could be applied to future scale-development projects.

Practical Applications in Academic Settings

The proven ability of AI systems to produce good scales to assess risks related to ChatGPT can find practical application right away. These validated instruments can be used in educational institutions to determine how students and faculty perceive the risks associated with AI, which can be informed in policy-making and training programs. The fact that the scales are multidimensional provides a subtle perception of a particular risk issue in various academic settings.

Also, the approach used in this paper provides a guideline to the creation of AI-guided measurement tools in other technology areas of emergence. With the infiltration of new technologies into academic settings, similarly designed methods would rapidly generate truthful measurement instruments to explain the perceptions and concerns held by stakeholders.

Limitations and Future Research

When evaluating the current results, the following limitations are to be considered. To begin with, a sample size of 120 respondents, although adequate in initial validation, limits the applicability of the results to larger academic groups. Follow-up research should be based on larger more representative cohorts in terms of a variety of institutional typologies and cultural milieus.

Second, the protocol of expert evaluation, which entailed the 20 credentialed raters, demonstrated a fair to moderate inter-rater reliability, which denoted some level of subjectivity in the item quality measurement.

Third, the possible biases in AI-generated items, which are not yet apparent in the first psychometric examination, should be further studied through more extensive research, through the qualitative analysis of the item content and linguistic nature. Future studies ought to challenge the idea that AI systems have systematic biases during the construction of items that may undermine measurement validity in diverse demographic population groups.

To sufficiently test the time-consistency of scales generated via AI, longitudinal research is required to determine the scale in different modes of administration. Also, comparative studies on several AI systems and human expert cohorts would improve the knowledge of the generalizability of the current results. The study of hybrid development approaches, which combine AI efficiency and human control, is also another relevant direction.

Ethical Considerations

Use of AI system in scale construction attracts critical ethical issues related to intellectual property, openness, and research integrity. Despite the high quality of AI-made items, it is still possible that doubts can remain about appropriate attribution and ethical aspects of research that need to be maintained by a human element. The future guidelines are expected to tackle these concerns and enable the beneficial application of AI in creating the measurement tools.

Conclusion

The current research provides solid empirical evidence that AI systems can be used to create psychometric scales that have a higher psychometric rigor than their human-created counterparts. The spectral reliability and validity assessments demonstrated that Claude scale achieved the best measures followed by ChatGPT; both were better than scales developed by experienced researchers in the methodology of research. The implication of these results is that a paradigm shift towards the development of AI-based or AI-enhanced scales as a new normative standard might be possible.

The implications of these implications, go beyond the efficiency gains, and bring up deeper issues of fundamental questions that bear on the nature of measurement expertise and the role of technology in developing scientific instruments. With the development of AI capabilities, the academic community should incorporate them with caution in the name of maintaining high standards of quality measurement and ethical research practices.

The verified tools aimed at the evaluation of perceived risks of ChatGPT in academic practice provide direct practical significance to higher-education institutions and researchers who explore the topic of technology usage in the field. In a broader sense, this research paper outlines an AI-aided scale development methodology that might streamline the development of measurement tools across the range of fields of research.

Future studies should seek to clarify the processes behind the perceived AI effectiveness in developing scales, setting up the optimal practice guidelines in human-AI cooperation in the designing of measurement scales, and extending the validation of AI-generated scales use in different populations and settings. The continuous transformation of psychometric studies under the influence of AI support creates never-before-seen prospects of developing the field of measurements, but at the cost of raising serious questions related to the changing status of research competency.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228-241. https://doi.org/10.1080/14703297.2023.2190148

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104-121. https://doi.org/10.1177/109442819800100106

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Publications.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Revelle, W. (2021). psych: Procedures for psychological, psychometric, and personality research [R package version 2.1.9]. Northwestern University. https://doi.org/10.32614/CRAN.package.psych

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric* *theory*. Routledge/Taylor & Francis Group.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838. https://doi.org/10.1177/0011000006288127