Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Psychometric evaluation of the UTAUT scale using the graded response model

Faridah Hanim Yahya [1,*], Aszunarni Ayob [2], Mohd Ridhuan Mohd Jamil [1], Abdussakir Abdussakir [3], Nurul Ain Mohd Daud [1]

[1]*Faculty of Human Development, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia*
[2]*Matriculation Division, Ministry of Education Malaysia, Complex E, 62604 Putrajaya, Malaysia*
[3]*Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia*

ARTICLE INFO

Article history:
Received 6 October 2024
Received in revised form
12 February 2025
Accepted 24 April 2025

Keywords:
Technology acceptance
Measurement validity
Psychometric analysis
Item response theory
Mathematics teachers

ABSTRACT

This study examined the reliability and validity of the Unified Theory of Acceptance and Use of Technology (UTAUT) measurement instrument. The sample included 202 mathematics teachers randomly selected from national secondary schools in Malaysia. The dataset was analyzed using the MIRT (Multidimensional Item Response Theory) and LTM (Latent Trait Models) packages in R software. The psychometric properties of the UTAUT scale were assessed using the graded response model (GRM), a type of Item Response Theory (IRT) model. The findings indicate that the scale effectively differentiates between various levels of technological acceptance, with most items showing high discrimination values. The threshold parameters suggest that higher response categories correspond to greater levels of agreement. The scale provides the highest accuracy in the middle range of traits but is less precise at the lower and upper extremes. However, the UTAUT scale still demonstrates good model fit and reliability.

1. Introduction

Since its introduction in 2003, the Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2012) has profoundly influenced technology acceptance research. Its ongoing relevance is demonstrated by its widespread application across diverse technological domains, from blockchain, e-government initiatives (Chen and Aklikokou, 2020; Li, 2021) and health-related technologies (Rouidi et al., 2022; Wang et al., 2020) to educational platforms such as mobile learning (Izkair and Lakulu, 2021; 2023) and e-learning system (Salloum and Shaalan, 2019) as well as digital learning tools (Abbad, 2021; Ustun et al., 2023; Wong et al., 2013; Yeop et al., 2019). Moreover, numerous studies utilizing UTAUT have exhibited enhanced predictive power for both behavioral intention and actual technology use in various settings. For instance, studies have demonstrated some ability to forecast technology adoption patterns among librarians in North America (Andrews et al., 2021), as well as robust predictive capabilities concerning healthcare workers' implementation of mobile electronic health record platforms (Kim et al., 2016).

With all that said, despite the model's widespread application, its psychometric evaluation over the years has largely relied on Classical Test Theory (CTT) approaches, such as Cronbach's alpha (Sezer and Yilmaz, 2019; Ustun et al., 2023), test-retest reliability (Sezer and Yilmaz, 2019; Ustun et al., 2023), and use of composite reliability in Partial Least Squares Structural Equation Modelling (PLS-SEM) as seen in the works by Ahmed et al. (2021) and Venkatesh et al. (2012). As the model's application expands to diverse cultural and linguistic contexts, an alternative psychometric approach becomes necessary. In this regard, Item Response Theory (IRT), specifically Samejima's (1969) Graded Response Model (GRM) (Samejima, 1969), emerges as a valuable alternative. Originally designed for analyzing ordered polytomous categories like Likert scales, GRM offers more detailed insights into item discrimination and difficulty parameters (Hambleton, 2021).

The widespread adoption of UTAUT across various technological domains has highlighted the need for a more robust evaluation of its measurement scale. Current psychometric

* Corresponding Author.
Email Address: faridahhanim@fpm.upsi.edu.my (F. H. Yahya)
https://doi.org/10.21833/ijaas.2025.04.017
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-0972-473X

assessments of UTAUT rely heavily on CTT approaches, which have known limitations, including sample dependency and the assumption of equal contribution of all items to the total score (Hambleton and Jones, 1993). As a result, these limitations may hinder CTT's ability to deliver the precise measurements needed for discerning nuanced differences in technology acceptance among diverse user groups or settings, which could potentially compromise the model's predictive and practical capabilities. To address these issues, this study seeks to fill this crucial gap by applying Samejima's (1969) GRM, an IRT approach, to evaluate the psychometric properties of the UTAUT scale. The application of IRT to UTAUT represents a significant methodological advancement, offering potential insights into item-level characteristics and measurement precision across the latent trait continuum - aspects that traditional CTT approaches fail to adequately capture. Research questions are as follows:

Q1: What is the overall fit of the UTAUT scale to the GRM, and how do the model parameters (e.g., item difficulty, discrimination) reflect the psychometric properties of the scale?
Q2: Does the UTAUT scale provide reliable measurement across the full range of the latent traits (theta), as evidenced by the test information function in the GRM?

2. Literature review

2.1. Brief introduction to UTAUT

Overall, the UTAUT emerged as a response to the fragmented landscape of technology acceptance models through a groundbreaking effort by Venkatesh et al. (2003). At its very core, Venkatesh's et al. (2003) UTAUT model proposed four fundamental constructs, namely Effort Expectancy, Performance Expectancy, Social Influence, and Facilitating Conditions.

As for its application, the enduring relevance of UTAUT is apparent. Various studies from all different kinds of sectors have successfully applied the model to study technology acceptance. In e-government, Chen and Aklikokou (2020) comprehensively studied in the Togolese context of 482 respondents, revealed that behavioral intention toward e-government services is primarily driven by perceived usefulness and ease of use, with these factors mediating between social influence, trustworthiness, and facilitating conditions.

Similarly, in the healthcare domain, UTAUT has guided digital health innovation implementation, with studies revealing practical insights for optimizing electronic health records and telemedicine platforms (Rouidi et al., 2022; Wang et al., 2020). Notably, Byrd et al.'s (2021) analysis of 1,254 healthcare providers demonstrated that perceived peer usage and social context explained 61-72% of implementation success, leading institutions to prioritize expanded access policies and early adopter visibility in their technology deployment strategies.

In higher education, UTAUT applications have yielded actionable insights for educational technology deployment. Xue's et al. (2024) systematic review of 159 studies revealed that performance expectancy (74%) and effort expectancy (50%) were the strongest drivers of technology adoption, shaping implementation strategies for faculty training and e-learning support services. These behavioral factors proved particularly crucial for VR adoption in STEM education, m-learning acceptance, and e-learning in education. Additionally, these insights have also become essential post-COVID-19 as universities enhance their digital learning ecosystems by prioritizing performance benefits and user support.

However, despite its widespread application, UTAUT has not been without criticism. Contended that UTAUT, in its attempt to be comprehensive, became unwieldy and lost parsimony mainly because it incorporated numerous variables and moderating factors, potentially obscuring the core determinants of technology acceptance and complicating the practical application and interpretation of the model (Marangunić and Granić, 2015; Tamilmani et al., 2021).

2.2. Methodological approaches in assessing reliability of UTAUT's constructs

2.2.1. Classical test theory approaches

For nearly two decades, UTAUT's psychometric properties have primarily been assessed using traditional reliability methods, particularly Cronbach's alpha within CTT. This trend, initiated by Venkatesh et al. (2003) who reported high internal consistency (Cronbach's alpha > 0.70) for UTAUT constructs, continues to influence recent research. For instance, a meta-analysis by Dwivedi et al. (2011), examining 18 studies, confirmed consistently high Cronbach's alpha values across UTAUT constructs. The average reliability coefficients were found to be Performance Expectancy (0.798), Effort Expectancy (0.870), Social Influence (0.811), Facilitating Conditions (0.747), Behavioral Intention (0.895), and Use Behavior (0.870).

While these findings appear robust, the persistent use of Cronbach's alpha in UTAUT research raises concerns about potential reliability inflation and the need for more sophisticated psychometric approaches. Cronbach's alpha has well-documented limitations when applied to complex, multidimensional constructs (McNeish, 2018; Sijtsma, 2009), particularly for the multifaceted constructs common in technology acceptance models. These limitations include:

1. Assumption of essential tau-equivalence: Cronbach's alpha relies on the premise that all

items in a scale are essentially tau-equivalent, meaning they contribute similarly (though not necessarily identically) to measuring the construct. However, this assumption is frequently violated, especially in complex, multidimensional measures. Raykov (1997) demonstrated that even when only one item in a scale deviates from essential tau-equivalence, it can significantly impact alpha's effectiveness as a reliability indicator. This violation often results in alpha underestimating the true reliability of a measure.

2. Sensitivity to the number of items: Cortina (1993) initially demonstrated alpha coefficient inflation with increased items, even amid low inter-item correlations, a phenomenon persistently overlooked in contemporary research (Dunn et al., 2014). Recent studies have corroborated this sensitivity, showing artificial inflation in long scales despite low inter-item correlations (Gu et al., 2013), and demonstrating how this leads to reliability overestimation (Trizano-Hermosilla and Alvarado, 2016).
3. Assumption of unidimensionality: Cronbach's alpha, while widely used, assumes unidimensionality in scale measurement, which may not be appropriate for the multidimensional constructs present in UTAUT. Green and Yang (2009) cautioned against the potential misuse of alpha in such multifaceted scales. Further research by Reise et al. (2013) offered empirical support for this concern, demonstrating that alpha can yield unreliable estimates for multidimensional constructs.

While Cronbach's alpha remains a common reliability measure, UTAUT researchers have also employed alternative CTT approaches, including test-retest reliability. As an example, Ustun et al. (2023) utilized test-retest reliability in their augmented reality UTAUT scale study, reporting a value of 0.97. However, it's important to note that test-retest reliability has limitations, particularly in the context of rapidly evolving technology. It assumes trait stability over time and can be susceptible to practice effects, potentially leading to overestimated reliability in dynamic technological environments (Ployhart and Vandenberg, 2009; Polit, 2014).

A notable evolution in UTAUT studies involves the integration of composite reliability, especially within PLS-SEM frameworks. Hair et al. (2017) championed this methodology in UTAUT research, contending that it provides a more precise evaluation of internal consistency for complex, multi-faceted constructs. Venkatesh et al. (2012) applied this method in their long-term investigation of an expanded UTAUT framework, revealing composite reliability scores above 0.80 for all UTAUT components. Nevertheless, this approach assumes equal indicator weights, which may not always suit complex UTAUT constructs (Bollen and Lennox, 1991; Schuberth et al., 2023).

2.2.2. The potential of item response theory approach

IRT is a modern psychometric approach that models the relationship between an individual's response to an item and the level of the latent trait being measured. In contrast to CTT, which emphasizes overall test statistics, IRT offers detailed item-specific data and enables more accurate assessment across various levels of the latent trait being evaluated (Embretson and Reise, 2013). Despite its potential, the application of IRT in UTAUT research remains limited, representing a significant opportunity for advancing the psychometric assessment of UTAUT. IRT offers several advantages over CTT approaches, including the ability to provide detailed information about item-level properties and differential item functioning. Among the various IRT frameworks, Samejima's (1969) GRM stands out as particularly relevant for UTAUT research, as it specifically addresses ordered polytomous data such as the Likert-scale responses frequently employed in UTAUT questionnaires (Samejima, 1969). In his research, Toland (2013) explained that the model is particularly well-suited for analyzing UTAUT scales, providing a more refined insight into item performance across varying degrees of technology acceptance. Applying IRT, especially GRM, to UTAUT provides several advantages. It enables researchers to enhance scale quality and efficacy by revealing item performance across trait levels (Embretson and Reise, 2013). The sample-independent nature of IRT parameters potentially yields more generalizable outcomes and enables reliable cross-study comparisons (Hambleton, 2021; Hambleton and Jones, 1993; Hambleton et al., 1993). Additionally, IRT facilitates differential item functioning (DIF) detection, ensuring fairness in measurement across diverse populations (Teresi and Fleishman, 2007). These advantages make IRT a powerful tool for advancing UTAUT research, potentially leading to more nuanced and culturally sensitive models of technology acceptance.

3. Research method

3.1. Respondents

This study surveyed 202 mathematics teachers from Malaysia's national secondary schools. These teachers were selected at random from the entire population of Malaysian Mathematics teachers.

3.2. Data collection and instrument

This study collected data through a survey utilizing a UTAUT-based instrument. Four UTAUT constructs were examined: Performance Expectancy, Effort Expectancy, Social Influence, and Facilitating Conditions. Each construct comprised four items rated on a 5-point Likert scale.

3.3. Data analysis

3.3.1. GRM

In general, the GRM predicts the cumulative probability that a person will select a certain response or a higher one based on their underlying ability or trait. This probability is formulated as:

$$P(X_{ijk} = k|\theta_j, b_{ik}, a_i) = P(X_{ijk} \geq k|\theta_j, b_{ik}, a_i) - P(X_{ijk} \geq k + 1|\theta_j, b_{ik}, a_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta_j - b_{i,k+1})}}$$

where, $i = 1, ..., I$ and I is the number of items. $j = 1, ..., n$ and n is the number of persons. $k = 1, ..., K$, and K is the number of response categories. X_{ijk} is the response k to item i for person j. a_i is the discrimination parameter for item i. b_{ik} is the threshold for category k of item i.

The model includes thresholds that separate response options. For example, a threshold (b_{i5}) distinguishes scores of 5 and above from scores of 4 and below. Tables 1 and 2 present the cut-off values for the item difficulty parameter, and the cut-off values for the item discriminant parameter, respectively.

Table 1: Cut-off values for item difficulty parameter

Cut-off values	Interpretation
< -2.00	Very easy
-2.00 – -0.50	Easy
-0.50 – 0.5	Medium
0.50 – 2.00	Hard
>2.00	Very hard

Table 2: Cut-off values for item difficulty parameter and item discriminant parameter

Cut-off values	Interpretation
0	None
0.01 – 0.34	Very low
0.35 – 0.64	Low
0.65 – 1.34	Moderate
1.35 – 1.69	High
> 1.70	Very high
+ Infinity	Perfect

3.3.2. Item operational characteristic curves (OCC), and item response category characteristic curves (ICC)

OCCs and ICCs are used in IRT to assess how well test items measure traits like ability or knowledge. OCCs demonstrate how likely someone is to select a specific response level (or any higher level) as their ability or knowledge increases. For lower trait levels, the curves for easier responses rise swiftly, whereas the curves for harder responses increase more gradually at higher trait levels. The 0.5 probability point on the curve marks a threshold, indicating that the likelihood of selecting one response equals that of selecting the next. The steepness of the curve, which is determined by the discrimination parameter, reflects how effectively the item differentiates between individuals at various trait levels. Fig. 1 shows an example of OCC. ICCs provide probabilistic representations of correct responses across trait levels. Steeper curves indicate that the item does a better job of distinguishing between individuals with different trait levels. For items with multiple response options, the points where the curves intersect show where one response becomes more likely than the next, providing insights into how well the item works and how respondents behave. Fig. 2 shows an example of ICC.

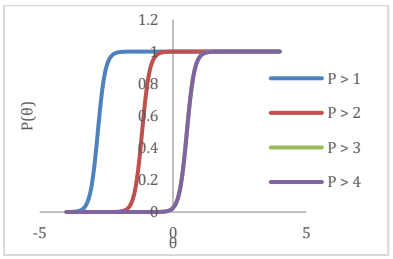

Fig. 1: An example of an OCC

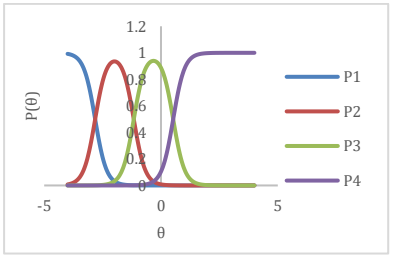

Fig. 2: An example of ICC

3.3.3. Assessing goodness of fit between two competing models

In evaluating the fit of two competing models, such as a simpler reduced model and a more complex full model, a Likelihood Ratio (LR) test can be employed. This test examines if adding more parameters improves the model's accuracy. Under the null hypothesis, the LR statistic follows a chi-squared (χ^2) distribution, with degrees of freedom equal to the parameter difference between the models. A higher LR value indicates that the Full Model is a better fit. We also use Akaike's (2011) Information Criterion (AIC) (Akaike, 2011) and Bayesian Information Criterion (BIC) (Neath and Cavanaugh, 2012) to assess model quality, with lower AIC and BIC values signaling the superior model.

3.3.4. Goodness of fit indices

Goodness-of-fit indices are essential for assessing how accurately a statistical model, like the GRM in IRT, represents the observed data structure. This study employed five key indices: Chi-Square (χ^2) Statistic, Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMSR), Tucker-Lewis Index (TLI), and Comparative Fit Index (CFI).

1. Chi-square statistic evaluates the difference between observed and expected covariance matrices, testing the null hypothesis of perfect model fit. A lower χ^2 value relative to degrees of freedom (df) suggests a better fit. However, this

test is sample size sensitive; large samples can make minor discrepancies statistically significant. Thus, while a non-significant result ($p > 0.05$) is desirable, it's not always the sole indicator of a good fit (Hair et al., 2019).

2. RMSEA estimates the average per-degree-of-freedom discrepancy between observed and predicted covariance matrices (Cook et al., 2009), adjusting for model complexity. Lower values indicate better fit: < 0.05 is considered good, 0.05-0.08 acceptable, and > 0.10 poor (Hair et al., 2019).
3. SRMSR, the standardized residuals measure, assesses the average discrepancy between observed and predicted correlations or covariances. Lower values indicate better fit, with < 0.08 typically considered good. It clearly shows how well the model accounts for data variance and covariance (Hair et al., 2019).
4. TLI, also known as the Non-Normed Fit Index (NNFI), compares the model's fit to a null model (assuming no variable relationships). It penalizes complexity and ranges from 0 to 1, with values closer to 1 indicating better fit. A TLI > 0.95 generally suggests a good fit. It adjusts for parameter count and model complexity.
5. CFI evaluates relative fit by comparing the model to a baseline (usually a null model with uncorrelated variables). It measures the proportion of fit improvement over the baseline. CFI ranges from 0 to 1, with values closer to 1 indicating better fit. A CFI > 0.90 is typically considered good. It's less sensitive to sample size than the Chi-Square test (Hair et al., 2019).

Table 3 presents the cut-off values for these fit indicators.

3.3.5. Test information function (TIF) and conditional standard error of measurement (CSEM)

The TIF and CSEM are important measures for checking the accuracy and reliability of a test across different levels of the trait being measured. TIF shows how effectively the test items collectively capture information about the trait, with higher TIF values suggesting greater accuracy in estimating trait levels. CSEM complements TIF by showing the precision of these estimates, with lower CSEM values indicating more reliable measurements. Both TIF and CSEM measures exhibit an inverse relationship: as TIF goes up, CSEM goes down, together offering a comprehensive understanding of the tests' reliability and accuracy. Fig. 3 shows an example of TIF.

Table 3: Cut-off values for goodness-of-fit indices

Indices	$m > 30$
Chi-square	Significant p-value is expected
SMSR	≤ 0.07 (with CFI > 0.92)
RMSEA	< 0.08 (with CFI > 0.92)
CFI	> 0.90
TLI	> 0.90

m: number of observed items

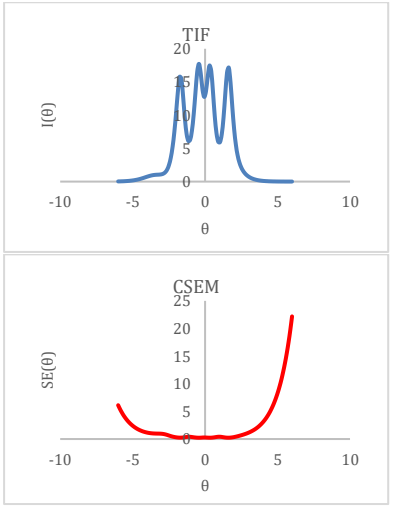

Fig. 3: An example of TIF and CSEM plot

3.3.6. Marginal reliability and empirical reliability

Marginal reliability measures how consistently an IRT model estimates a person's ability across different levels of ability in a group. It is akin to Cronbach's alpha but designed for IRT models. This reliability is calculated by comparing the estimated ability scores (true variance) to measurement errors. A higher marginal reliability means the model gives more accurate and consistent ability estimates.

On the other hand, empirical reliability checks how closely the observed test scores match the true ability scores in a dataset, often using methods like Expected a posteriori (EAP). High empirical reliability suggests that the observed scores are good representations of actual abilities. Both reliability measures have a range of 0 to 1, with higher values suggesting more accurate scores with fewer errors.

Table 4 provides the cut-off values for interpreting these reliability parameters, offering guidelines for assessing the quality of measurement in IRT models.

3.3.7. Software package for GRM

The software packages used to implement the GRM were mirt and ltm, both available in the R programming environment.

4. Results and discussions

Two graded response models were fitted to UTAUT scale data. These two models are the Reduced Model which assumes that the discriminant parameters, a_i, are constant, and the Full Model which allows the discriminant parameters, a_i, to vary. Specifically, in the Reduced Model, all items are believed to give the same amount of information about respondents' attitudes toward the importance of the UTAUT. On the contrary, in the Full Model, all items are assumed to provide different amounts of information about respondents' attitudes. Table 5 shows the findings of the LR tests for all constructs

of the UTAUT scale. Based on Table 5, the *p*-values are significant for all constructs. Hence, the Reduced Models are rejected. The next discussions on the items are based on the Full Models.

Table 4: Cut-off values for item difficulty parameter

Cut-off values		Interpretation
Below 0.60	Poor reliability	The model's estimates are not reliable, with a significant amount of measurement error.
0.60 to 0.69	Marginal reliability	The model's estimates have a noticeable amount of error and may not be reliable enough for certain applications.
0.70 to 0.79	Acceptable reliability	The estimates are somewhat reliable but there is some measurement error.
0.80 to 0.89	Good reliability	The model's estimates are reliable for most practical purposes.
0.90 and above	Excellent reliability	The model provides highly consistent estimates of the latent trait across different levels of the trait.

Table 5: Result of likelihood ratio (LR) test for the UTAUT scale

Construct	Model	AIC	BIC	Log likelihood	LR test	df	Decision
Performance expectancy	Reduced model	1500.3	1550.0	−735.17	-	-	Full model is better
	Full model	1419.8	1479.3	−691.89	86.55*	3	
Effort expectancy	Reduced model	1367.5	1423.7	−666.74	-	-	Full model is better
	Full model	1258.7	1324.9	−609.34	114.79*	3	
Social influence	Reduced model	1390.6	1453.4	−676.28	-	-	Full model is better
	Full model	1330.2	1396.4	−645.09	62.38*	1	
Facilitating conditions	Reduced model	1875.8	1932.1	−920.92	-	-	Full model is better
	Full model	1772.3	1838.5	−866.16	109.52*	3	

*: p <0.001

Table 6 shows the item difficulty category (b_{ik}) and the value of the discriminant parameter (a_i) for the Full Model. The last row indicates the value that separates each response category. Since items are based on a 5-point Likert scale, there are five thresholds that separate each of these response categories. Since each respondent has a 100% chance to choose the "Strongly disagree" response category, there is no threshold for the response category.

The values of the discriminant parameter (a_i) for all items are positive. Except for item JP4, all items have very high discriminant values, indicating that these items can discriminate between respondents with low and high trait levels. Quite the reverse, the discriminant value for item JP4 is considered moderate, implying that this item is not good in discriminating between respondents with low and high trait levels.

The thresholds for all items span from negative to positive sections of the trait, ascendingly. In other words, the higher response category has higher item locations, indicating endorsement of more UTAUT scale.

Table 6: UTAUT scale parameter estimates using a GRM

Construct	Item code	Item description	B_1	B_2	B_3	B_4	A
Performance expectancy	JP1	I think the 3D geometry pedagogical module based on augmented reality will be useful in my teaching.	-	−2.82	−1.17	0.52	4.11 (Very high)
	JP1	I think the 3D geometry pedagogical module based on augmented reality will be useful in my teaching.	-	−2.82	−1.17	0.52	4.11 (Very high)
	JP2	The use of the 3D geometry pedagogical module based on augmented reality will allow me to complete teaching tasks more quickly.	-	−2.07	−1.08	0.71	3.46 (Very high)
	JP3	The use of the 3D geometry pedagogical module based on augmented reality will increase my work productivity.	−2.58	−1.92	−1.01	0.53	31.47 (Very high)
	JP4	If I use the 3D geometry pedagogical module based on augmented reality, the skills gained from this module will increase my chances of getting an outstanding service award.	−3.04	−1.23	−0.14	1.75	1.20 (Moderate)
Effort expectancy	JU5	The use of the 3D geometry pedagogical module based on augmented reality will be clear and easy to understand.	−2.82	−1.84	−0.69	0.81	5.18 (Very high)
	JU6	I will easily master the 3D geometry pedagogical module based on augmented reality.	−2.83	−1.75	−0.65	0.99	4.71 (Very high)
	JU7	I think the 3D geometry pedagogical module based on augmented reality will be easy to use.	−2.86	−1.88	−0.47	1.07	4.63 (Very high)
	JU8	I think learning the use of the 3D geometry pedagogical module based on augmented reality will be easy for me.	−2.60	−1.85	−0.74	1.040	3.38 (Very high)
Social influence	PS9	Individuals who influence my behavior think I need to use the 3D Geometry Pedagogical Module based on Augmented Reality.	−2.83	−1.44	−0.36	1.18	4.76 (Very high)
	PS 10	My colleagues encouraged me to use the pedagogical module of the 3D geometry module based on augmented reality.	−2.70	−1.43	−0.41	1.18	10. 84 (Very high)
	PS 11	The researchers who carried out this study assisted in the continuous use of the pedagogical module of the 3D geometry pedagogical module based on augmented reality.	−2.56	−1.55	−0.50	1.07	3.70 (Very high)
	PS 12	The school provides support for the use of the 3D geometry pedagogical module based on augmented reality.	−3.00	-1.55	−0.33	1.43	3.22 (Very high)
Facilitating conditions	KP 13	I have the necessary equipment to use the 3D Geometry Pedagogical Module based on Augmented Reality.	−1.90	−0.55	0.24	1.62	4.31 (Very high)
	KP 14	I have knowledge of the procedures for using the 3D geometry pedagogical module based on augmented reality.	−1.68	−0.43	0.36	1.60	6.22 (Very high)
	KP 15	I feel that the 3D geometry pedagogical module based on augmented reality cannot be adapted to other teaching methods that I use (reverse-coded).	−3.46	−0.48	0.47	2.12	1.95 (Very high)
	KP 16	I have someone to refer to if I encounter problems related to the use of the 3D geometry pedagogical module based on augmented reality.	−1.59	−0.50	0.44	1.74	2.84 (Very high)
			1 vs 2–5	1–2 vs 3–5	1–3 vs 4–5	1–4 vs 5	

B_1, B_2, B_3, and B_4 are the threshold parameters for each item

Fig. 4 presents the item operating characteristic curves (OCCs) for all items, showing the probability of selecting each category or higher as a function of latent attitude. Thresholds are indicated by intersections with the 0.5 probability lines. Each curve has a distinct slope at this line, visible in its OCC. Items with high discrimination values (like JP3 and PS10) demonstrate steeper slopes than those with lower values, reflecting their greater ability to differentiate between respondents at various levels of the measured trait.

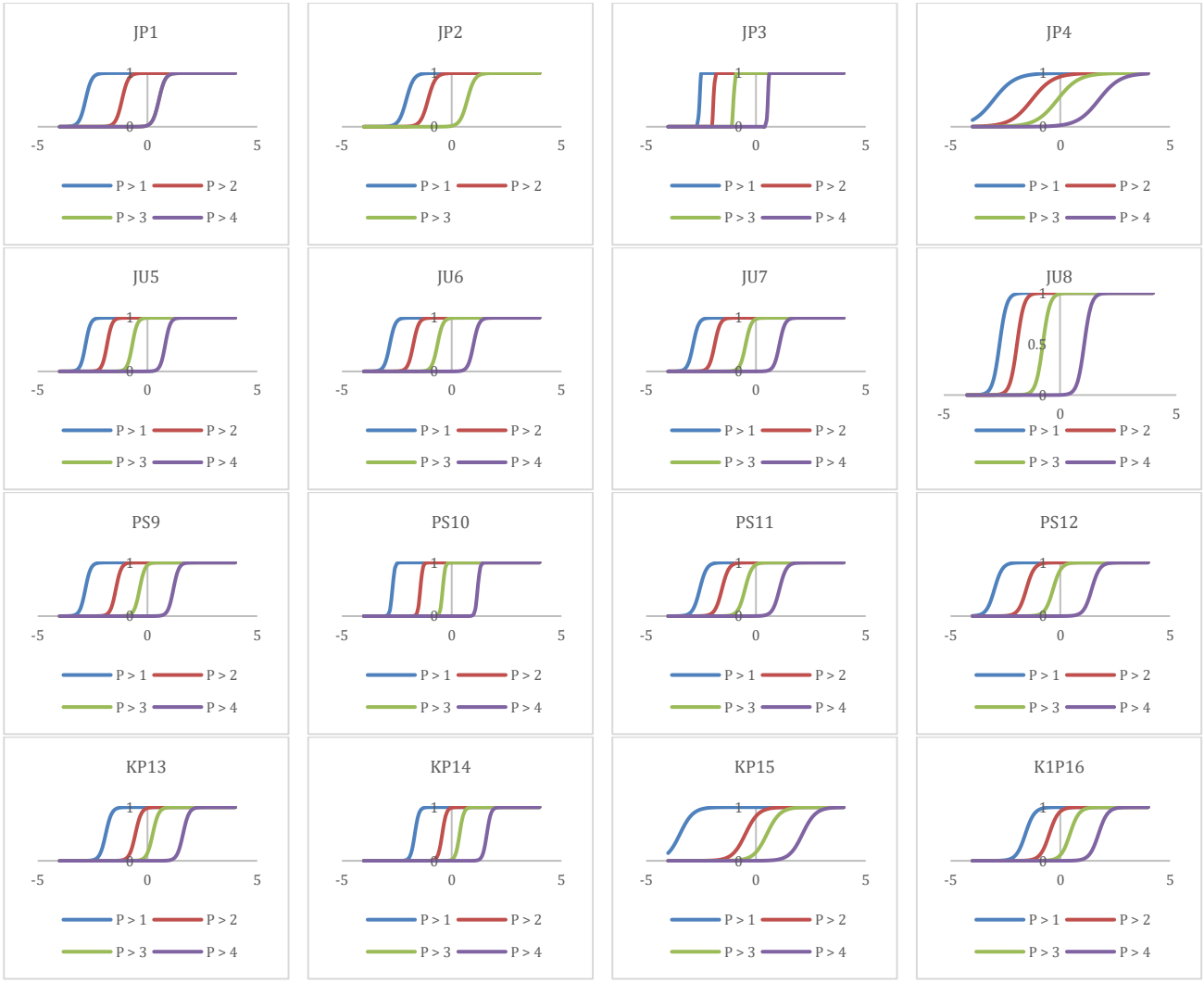

Fig. 4: OCC for all items

Fig. 5 displays the ICCs. The ICC curves, which represent the relative position of each category along the trait levels, can be used to evaluate response categories for each item. Except for item JP3, each response category for all items is most likely to be selected at different trait levels, as depicted in the graphs. Specifically, category 5 is the most dominant for all items. On the contrary, the response categories for item JP4 were not monotonically related to performance expectancy, and there was somewhat vague evidence regarding the selection of each response category along the trait levels. Fig. 6 shows the graphs of TIF and CSEM for all constructs. In general, the information initially increases as the trait level rises, reaching a peak where the test is most informative. Beyond this peak, the information starts to decline, indicating that the test becomes less precise at very high or very low levels of the latent trait. On the contrary, the inverse relationship between TIF and CSEM means that as the information increases, the CSEM decreases, reflecting higher measurement precision.

Precisely, for Performance Expectancy and Facilitating Conditions constructs, the tests provide high information in the middle ranges (−3 < theta <3). This indicates that at these points, the tests can distinguish between individuals with different levels of the trait with high accuracy. On the other hand, the tests provide low information at both extremes, suggesting that the tests might be less reliable or informative at these extreme ends of the latent trait continuum. For Effort Expectancy and Social Influence constructs, the tests provide high information in the greater middle range (−4 < theta <3), and similarly low information at both extremes. As depicted in the graphs, the CSEM values were at the lowest positions when the TIF values were at the highest ranges, and vice versa. This reduced accuracy at the extreme ends of the ability scale stems from individuals with very low or high trait

levels consistently providing similar responses, which impedes accurate assessment. Therefore, to improve measurements at these ends of the scale, experts recommend incorporating items with broader difficulty levels and implementing adaptive testing procedures that adjust to each individual's trait levels (Samejima, 1969). Table 7 displays model fit statistics for four key constructs in a structural equation framework: Performance Expectancy, Effort Expectancy, Social Influence, and Facilitating Conditions. Table 7 provides a comprehensive assessment of each construct's fit using multiple indicators, including Chi-Square, degrees of freedom (df), p-value RMSEA, SRMSR, TLI, and CFI.

Table 7: Goodness-of-fit indices

Construct	Chi-square	*Df*	*P*	RMSEA	SMSR	TLI	CFI
Performance expectancy	5.49	2	0.06	0.07	0.04	0.98	0.99
Effort expectancy	4.32	2	0.12	0.07	0.03	0.99	0.99
Social influence	4.24	2	0.25	0.06	0.04	0.90	0.97
Facilitating conditions	1.83	2	0.40	0.00	0.04	1.000	1.00

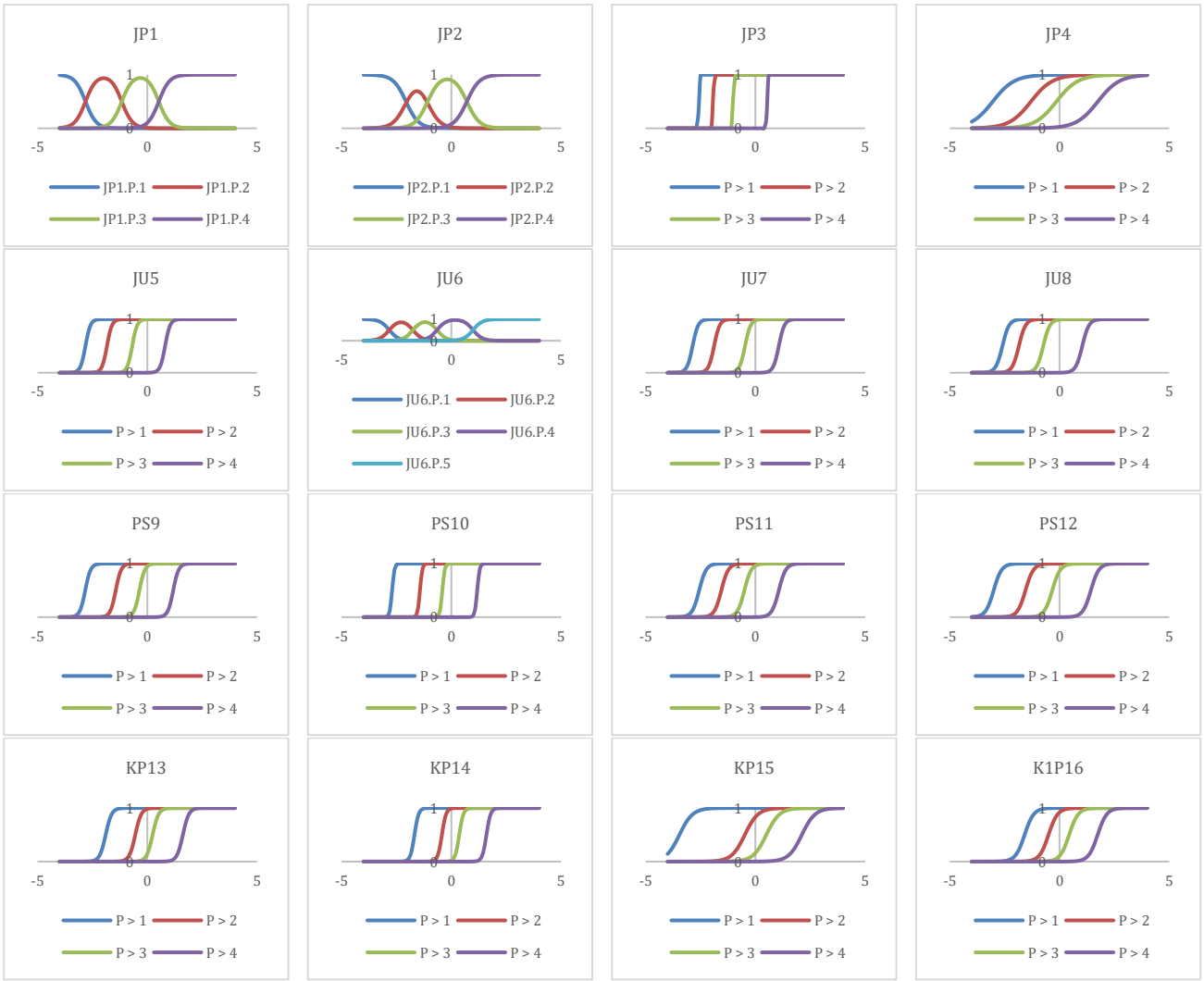

Fig. 5: ICC for all items

Performance Expectancy demonstrates an acceptable fit, with a Chi-Square value of 5.488 (df = 2, p = 0.064) slightly exceeding the 0.05 threshold. Its RMSEA of 0.073 falls within the preferred range, while the SRMSR (0.042), TLI (0.976), and CFI (0.992) all indicate a good fit. Effort Expectancy exhibits a stronger fit, boasting a Chi-Square value of 4.316 (df = 2, p = 0.116) well above the 0.05 threshold. The construct's RMSEA (0.076) is acceptable, and its SRMSR (0.027), TLI (0.989), and CFI (0.996) approach ideal values suggest an excellent model fit. For Social Influence, the Chi-Square value of 4.235 (df = 2, p = 0.254) indicates a good fit. Its RMSEA (0.065) is acceptable, while the SRMSR (0.043), TLI (0.902), and CFI (0.967) suggest a reasonable fit overall.

Facilitating Conditions stands out with exceptional fit indices. Its Chi-Square value of 1.831 (df = 2, p = 0.4) far exceeds the threshold, complemented by a perfect RMSEA of 0.00. The construct's SRMSR (0.041) and perfect TLI and CFI (both 1.000) further confirm its excellent fit.

In summary, Effort Expectancy and Facilitating Conditions showcase excellent model fit, while Performance Expectancy presents an acceptable fit with minor concerns. Social Influence demonstrates

a reasonable fit based on the analyzed indices. Given these results, all items will be retained for the UTAUT scale, as presented in Table 8. The reported values indicate robust reliability for the evaluated items, with both types of reliability provided.

5. Conclusion

This study applied Samejima's (1969) GRM to evaluate the psychometric properties of the UTAUT scale. The findings reveal that the discriminant parameters for all items are positive, with most items having high discriminant values, meaning they effectively distinguish between individuals with different levels of the underlying trait. The threshold parameters progress from negative to positive, indicating that higher response categories correspond to higher levels of agreement with the UTAUT scale items. These thresholds cover a wide range of trait continuums, reflecting varied levels of respondent agreement. To understand what these threshold parameters mean in practice, we examined measurement precision across the trait continuum. Our analysis shows that the UTAUT scale works best in measuring technology acceptance in the middle range, where most respondents fall. However, the scale becomes less precise at the extremes, where respondents show either very low or very high acceptance levels, primarily due to these individuals consistently selecting extreme responses. This limitation could be addressed through broader difficulty ranges in items and adaptive testing procedures (Samejima, 1969).

In conclusion, despite the identified limitations in measuring extreme responses, the UTAUT model demonstrates robust psychometric properties with acceptable model fit and high reliability. These findings substantiate the overall psychometric integrity of the UTAUT scale, validating its utility as a measurement instrument for technology acceptance assessment.

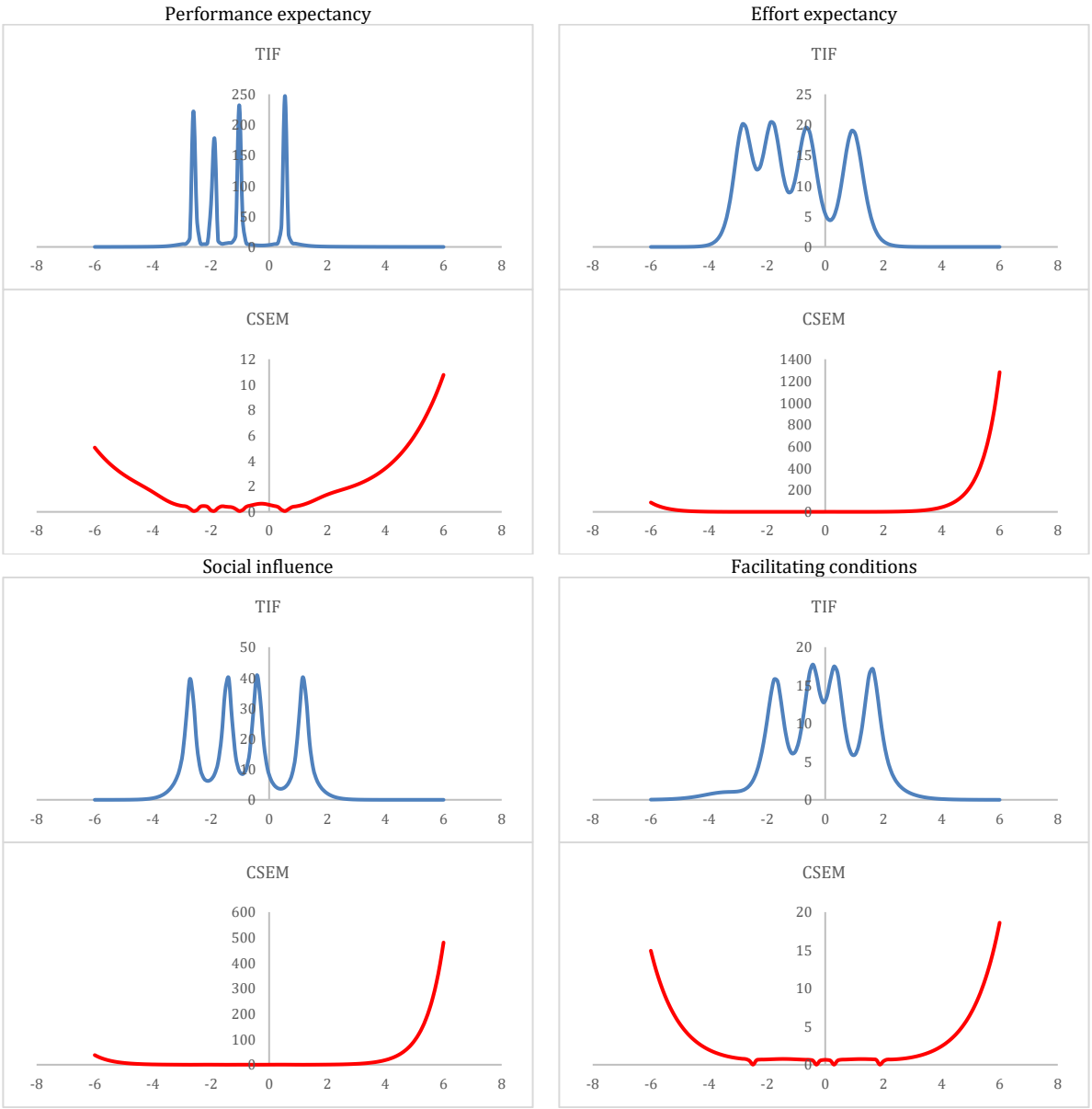

Fig. 6: Test information functions and conditional standard error

Table 8: Final items for UTAUT scale, marginal reliability, and empirical reliability

Construct	Item code	Item description	Marginal reliability	Empirical reliability
Performance Expectancy	JP1	I think the 3D geometry pedagogical module based on augmented reality will be useful in my teaching.	0.808	0.861
	JP2	The use of the 3D geometry pedagogical module based on augmented reality will allow me to complete teaching tasks more quickly.		
	JP3	The use of the 3D geometry pedagogical module based on augmented reality will increase my work productivity.		
	JP4	If I use of the 3D geometry pedagogical module based on augmented reality, the skills gained from this module will increase my chances of getting an outstanding service award.		
Effort Expectancy	JU5	The use of the 3D geometry pedagogical module based on augmented reality will be clear and easy to understand.	0.885	0.893
	JU6	I will easily master the 3D geometry pedagogical module based on augmented reality.		
	JU7	I think the 3D geometry pedagogical module based on augmented reality will be easy to use.		
	JU8	I think learning the use of the 3D geometry pedagogical module based on augmented reality will be easy for me.		
Social Influence	PS9	Individuals who influence my behavior think I need to use the 3D geometry pedagogical module based on augmented reality.	0.891	0.909
	PS10	Colleagues encouraged me to use the pedagogical module of the 3D geometry module based on augmented reality.		
	PS11	The researchers who carried out this study assisted in the continuous use of the pedagogical module of the 3D geometry pedagogical module based on augmented reality.		
	PS12	The school provides support for the use of the 3D geometry pedagogical module based on augmented reality.		
Facilitating Conditions	KP13	I have the necessary equipment to use the 3D geometry pedagogical module based on augmented reality.	0.910	0.913
	KP14	I have knowledge of the procedures for using the 3D geometry pedagogical module based on augmented reality.		
	KP15	I feel that the 3D geometry pedagogical module based on augmented reality cannot be adapted to other teaching methods that I use (reverse-coded).		
	KP16	I have someone to refer to if I encounter problems related to the use of the 3D geometry pedagogical module based on augmented reality.		

5.1. Implications of the study

The implications of this study have broad relevance for those using the UTAUT scale in research and practice. By applying Samejima's (1969) GRM, the study has provided detailed insights into the psychometric properties of the UTAUT scale, confirming its effectiveness in distinguishing different levels of technology acceptance and providing reliable insights into user behavior. In addition, its broad threshold parameters allow for comprehensive assessment across various trait levels. For practitioners, particularly those involved in the implementation and promotion of new technologies, the study's findings suggest that the UTAUT scale is a valuable tool for identifying potential challenges or resistance points in user adoption. The scale's ability to provide detailed information within the middle range of the trait spectrum means it can effectively guide interventions and strategies aimed at improving technology uptake.

In addition, the UTAUT scale has significant practical implications in various domains like education and healthcare by enabling a more nuanced understanding of technology adoption. In educational settings, it helps institutions effectively evaluate and address factors affecting digital tool adoption, including e-learning platforms (Abbad, 2021), enabling evidence-based decisions about training programs, interface design, and accessibility improvements. In healthcare, the scale has proven valuable for understanding technology adoption patterns, from electronic health records to telemedicine platforms, by identifying key barriers related to usability, trust, and infrastructure (Byrd et al., 2021; Rouidi et al., 2022; Wang et al., 2020; Yu and Chen, 2024). Overall, the UTAUT scale supports the development of inclusive, scalable, and user-centered solutions, ensuring that technology effectively meets the diverse needs of users.

5.2. Limitations and recommendations for future research

This study highlights that while the UTAUT scale effectively distinguishes between different levels of technology acceptance, its accuracy diminishes at the extreme ends of the trait continuum. The scale is most precise in the middle range of technology acceptance but provides less reliable information for individuals at very high or very low levels. This limitation affects the overall comprehensiveness of the scale in capturing the full spectrum of technology acceptance.

Future research should explore methods to enhance the UTAUT scale's precision at the extremes of the technology acceptance spectrum. This could involve developing additional items or adjusting the scale to improve accuracy for individuals with very high or low levels of technological acceptance. Alternatively, implementing adaptive testing could be explored as a solution.

List of abbreviations

UTAUT	Unified theory of acceptance and use of technology
CTT	Classical test theory
IRT	Item response theory
GRM	Graded response model
PLS-SEM	Partial least squares structural equation modeling
DIF	Differential item functioning

OCC	Operational characteristic curve
ICC	Item response category characteristic curve
TIF	Test information function
CSEM	Conditional standard error of measurement
LR	Likelihood ratio
AIC	Akaike's information criterion
BIC	Bayesian information criterion
RMSEA	Root mean square error of approximation
SRMSR	Standardized root mean square residual
TLI	Tucker-Lewis index
CFI	Comparative fit index
EAP	Expected a posteriori
mirt	Multidimensional item response theory (R software package)
ltm	Latent trait models (R software package)
df	Degrees of freedom
A	Discrimination parameter in GRM
B1–B4	Threshold parameters for Likert scale responses
JP1–JP4	Item codes under performance expectancy construct
JU5–JU8	Item codes under effort expectancy construct
PS9–PS12	Item codes under social influence construct
KP13–KP16	Item codes under facilitating conditions construct

Acknowledgment

This research project was co-funded by the Universiti Pendidikan Sultan Idris, Malaysia under University Research Grant Special Interest Group (SIG) 2022 (Code: 2022-0028-106-01).

Compliance with ethical standards

Ethical considerations

This study upheld informed consent, confidentiality, and ethical compliance while protecting participants' rights and data security.

Conflict of interest

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

References

Abbad MMM (2021). Using the UTAUT model to understand students' usage of e-learning systems in developing countries. Education and Information Technologies, 26(6): 7205-7224. https://doi.org/10.1007/s10639-021-10573-5 **PMid:34025204 PMCid:PMC8122219**

Ahmed RR, Štreimikienė D, and Štreimikis J (2021). The extended UTAUT model and learning management system during COVID-19: Evidence from PLS-SEM and conditional process modeling. Journal of Business Economics and Management, 23(1): 82-104. https://doi.org/10.3846/jbem.2021.15664

Akaike H (2011). Akaike's information criterion. In: Lovric M (Ed.), International encyclopedia of statistical science: 25-25. Springer Berlin Heidelberg, Berlin, Germany. https://doi.org/10.1007/978-3-642-04898-2_110

Andrews JE, Ward H, and Yoon J (2021). UTAUT as a model for understanding intention to adopt AI and related technologies among librarians. The Journal of Academic Librarianship, 47(6): 102437. https://doi.org/10.1016/j.acalib.2021.102437

Bollen K and Lennox R (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110(2): 305-314. https://doi.org/10.1037//0033-2909.110.2.305

Byrd IV TF, Kim JS, Yeh C, Lee J, and O'Leary KJ (2021). Technology acceptance and critical mass: Development of a consolidated model to explain the actual use of mobile health care communication tools. Journal of Biomedical Informatics, 117: 103749. https://doi.org/10.1016/j.jbi.2021.103749 **PMid:33766780**

Chen L and Aklikokou AK (2020). Determinants of e-government adoption: Testing the mediating effects of perceived usefulness and perceived ease of use. International Journal of Public Administration, 43(10): 850-865. https://doi.org/10.1080/01900692.2019.1660989

Cook KF, Kallen MA, and Amtmann D (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Quality of Life Research, 18(4): 447-460. https://doi.org/10.1007/s11136-009-9464-4 **PMid:19294529 PMCid:PMC2746381**

Cortina JM (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78(1): 98-104. https://doi.org/10.1037//0021-9010.78.1.98

Dunn TJ, Baguley T, and Brunsden V (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3): 399-412. https://doi.org/10.1111/bjop.12046 **PMid:24844115**

Dwivedi YK, Rana NP, Chen H, and Williams MD (2011). A meta-analysis of the unified theory of acceptance and use of technology (UTAUT). In: Nüttgens M, Gadatsch A, Kautz K, Schirmer I, Blinn N (Eds), Governance and sustainability in information systems: Managing the transfer and diffusion of IT: 155-170. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-24148-2_10

Embretson SE and Reise SP (2013). Item response theory. Psychology Press, Hove, UK. https://doi.org/10.4324/9781410605269

Green SB and Yang Y (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. Psychometrika, 74(1): 155-167. https://doi.org/10.1007/s11336-008-9099-3

Gu F, Little TD, and Kingston NM (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. Methodology, 9(1): 30-40. https://doi.org/10.1027/1614-2241/a000052

Hair JF, Black WC, Babin BJ, and Anderson RE (2019). Multivariate data analysis. 8th Edition, Cengage Learning, Andover, UK.

Hair JF, Matthews LM, Matthews RL, and Sarstedt M (2017). PLS-SEM or CB-SEM: Updated guidelines on which method to use. International Journal of Multivariate Data Analysis, 1(2): 107-123. https://doi.org/10.1504/IJMDA.2017.087624

Hambleton RK (2021). Handbook of item response theory. Chapman and Hall, London, UK.

Hambleton RK and Jones RW (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12(3): 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton RK, Jones RW, and Rogers HJ (1993). Influence of item parameter estimation errors in test development. Journal of Educational Measurement, 30(2): 143-155. https://doi.org/10.1111/j.1745-3984.1993.tb01071.x

Izkair AS and Lakulu MM (2021). Experience moderator effect on the variables that influence intention to use mobile learning. Bulletin of Electrical Engineering and Informatics, 10(5): 2875-2883. https://doi.org/10.11591/eei.v10i5.3109

Izkair AS and Lakulu MM (2023). The moderating effects of gender on factors affecting the intention to use mobile learning. Journal of Information Technology Education: Research, 22: 199-233. https://doi.org/10.28945/5094

Kim S, Lee KH, Hwang H, and Yoo S (2016). Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital. BMC Medical Informatics and Decision Making, 16(1): 12. https://doi.org/10.1186/s12911-016-0249-8 **PMid:26831123 PMCid:PMC4736616**

Li W (2021). The role of trust and risk in citizens' e-government services adoption: A perspective of the extended UTAUT model. Sustainability, 13(14): 7671. https://doi.org/10.3390/su13147671

Marangunić N and Granić A (2015). Technology acceptance model: A literature review from 1986 to 2013. Universal Access in the Information Society, 14(1): 81-95. https://doi.org/10.1007/s10209-014-0348-1

McNeish D (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3): 412-433. https://doi.org/10.1037/met0000144 **PMid:28557467**

Neath AA and Cavanaugh JE (2012). The Bayesian information criterion: Background, derivation, and applications. WIREs Computational Statistics, 4(2): 199-203. https://doi.org/10.1002/wics.199

Ployhart RE and Vandenberg RJ (2009). Longitudinal research: The theory, design, and analysis of change. Journal of Management, 36(1): 94-120. https://doi.org/10.1177/0149206309352110

Polit DF (2014). Getting serious about test–retest reliability: A critique of retest research and some recommendations. Quality of Life Research, 23(6): 1713-1720. https://doi.org/10.1007/s11136-014-0632-9 **PMid:24504622**

Raykov T (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. Multivariate Behavioral Research, 32(4): 329-353. https://doi.org/10.1207/s15327906mbr3204_2 **PMid:26777071**

Reise SP, Bonifay WE, and Haviland MG (2013). Scoring and modeling psychological measures in the presence of multidimensionality. Journal of Personality Assessment, 95(2): 129-140. https://doi.org/10.1080/00223891.2012.725437 **PMid:23030794**

Rouidi M, Elouadi AE, Hamdoune A, Choujtani K, and Chati A (2022). TAM-UTAUT and the acceptance of remote healthcare technologies by healthcare professionals: A systematic review. Informatics in Medicine Unlocked, 32: 101008. https://doi.org/10.1016/j.imu.2022.101008

Salloum SA and Shaalan K (2019). Factors affecting students' acceptance of e-learning system in higher education using UTAUT and structural equation modeling approaches. In: Hassanien A, Tolba M, Shaalan K, & Azar A (Eds), Proceedings of the international conference on advanced intelligent systems and informatics 2018: 469–480. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-99010-1_43

Samejima F (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34(4, Pt 2): 100-100. https://doi.org/10.1007/BF03372160

Schuberth F, Zaza S, and Henseler J (2023). Partial least squares is an estimator for structural equation models: A comment on Evermann and Rönkkö (2021). Communications of the Association for Information Systems, 52(1): 711-714. https://doi.org/10.17705/1CAIS.05232

Sezer B and Yilmaz R (2019). Learning management system acceptance scale (LMSAS): A validity and reliability study. Australasian Journal of Educational Technology, 35(3): 15-30. https://doi.org/10.14742/ajet.3959

Sijtsma K (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1): 107-120. https://doi.org/10.1007/s11336-008-9101-0 **PMid:20037639 PMCid:PMC2792363**

Tamilmani K, Rana NP, and Dwivedi YK (2021). Consumer acceptance and use of information technology: A meta-analytic evaluation of UTAUT2. Information Systems Frontiers, 23(4): 987-1005. https://doi.org/10.1007/s10796-020-10007-6

Teresi JA and Fleishman JA (2007). Differential item functioning and health assessment. Quality of Life Research, 16(1): 33-42. https://doi.org/10.1007/s11136-007-9184-6 **PMid:17443420**

Toland MD (2013). Practical guide to conducting an item response theory analysis. The Journal of Early Adolescence, 34(1): 120-151. https://doi.org/10.1177/0272431613511332

Trizano-Hermosilla I and Alvarado JM (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. Frontiers in Psychology, 7: 769. https://doi.org/10.3389/fpsyg.2016.00769 **PMid:27303333 PMCid:PMC4880791**

Ustun AB, Karaoglan-Yilmaz FG, and Yilmaz R (2023). Educational UTAUT-based virtual reality acceptance scale: A validity and reliability study. Virtual Reality, 27(2): 1063-1076. https://doi.org/10.1007/s10055-022-00717-4

Venkatesh V, Morris MG, Davis GB, and Davis FD (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3): 425-478. https://doi.org/10.2307/30036540

Venkatesh V, Thong JYL, and Xu X (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. MIS Quarterly, 36(1): 157-178. https://doi.org/10.2307/41410412

Wang H, Tao D, Yu N, and Qu X (2020). Understanding consumer acceptance of healthcare wearable devices: An integrated model of UTAUT and TTF. International Journal of Medical Informatics, 139: 104156. https://doi.org/10.1016/j.ijmedinf.2020.104156 **PMid:32387819**

Wong KT, Teo T, and Russo S (2013). Interactive whiteboard acceptance: Applicability of the UTAUT model to student teachers. The Asia-Pacific Education Researcher, 22: 1-10. https://doi.org/10.1007/s40299-012-0001-9

Xue L, Rashid AM, and Ouyang S (2024). The unified theory of acceptance and use of technology (UTAUT) in higher education: A systematic review. Sage Open, 14(1). https://doi.org/10.1177/21582440241229570

Yeop MA, Yaakob MFMW, Wong KT, Don Y, and Zain FM (2019). Implementation of ICT policy (blended learning approach): Investigating factors of behavioural intention and use behaviour. International Journal of Instruction, 12(1): 767-782. https://doi.org/10.29333/iji.2019.12149a

Yu S and Chen T (2024). Understanding older adults' acceptance of Chatbots in healthcare delivery: An extended UTAUT model. Frontiers in Public Health, 12: 1435329. https://doi.org/10.3389/fpubh.2024.1435329 **PMid:39628811 PMCid:PMC11611720**