

Development and Initial Validation of a Survey Instrument on Students' Acceptance of and Motivation Toward Generative AI in Higher Education

Ng Jun Wei[1,2*]

[1]Lecturer, School of Health and Life Sciences, PSB Academy, Singapore.
[2]Doctoral Candidate, Universidad Católica San Antonio de Murcia, Spain.
*Corresponding Author Email: pt.ngjw@gmail.com

DOI: 10.83834/IJTHS.0204060 *(Received: 19.10.2026; Accepted: 24.11.2026)*

Abstract

Questionnaires on students' acceptance of generative AI (GenAI) are multiplying, but few combine technology acceptance with the expectancy-value beliefs and psychological needs that shape learning, and few test whether scores predict what students actually do. This study reports the development and initial validation of the 45-item Values, Expectancy and Motivation Perception of AI in Education (VEM-AIEd) questionnaire, which has 15 three-item scales drawn from the Unified Theory of Acceptance and Use of Technology, expectancy-value theory and self-determination theory, together with a scale on ethical considerations. Eighteen items, covering six scales, were negatively worded. Sport science students at a private higher education institution in Singapore (N = 140) completed the questionnaire in 2024. The designed 15-factor model fitted moderately (CFI = 0.868, RMSEA = 0.062, SRMR = 0.071) and far better than one- and two-factor alternatives. Thirteen scales showed good internal consistency (α = 0.78–0.87) and convergent validity (AVE ≥ 0.50); social influence and cost did not. Item wording, however, shaped the structure. Exploratory factor analysis separated negatively from positively worded items, and latent correlations averaged 0.53 among negatively worded scales but only 0.11 between negatively and positively worded scales, even where content overlapped (performance expectancy and utility value, r = 0.03). Six-week test–retest reliability was poor to moderate (ICC = 0.20–0.55), and no scale predicted students' logged use of a course chatbot. A shorter, consistently worded version with seven scales and 21 items fitted well (CFI = 0.939) but requires confirmation in new samples. The findings caution against mixing item directions in GenAI acceptance instruments.

Keywords: *Generative AI; Questionnaire validation; Confirmatory factor analysis; Technology acceptance; Expectancy-value theory; Self-determination theory; Negatively worded items; Higher education*

INTRODUCTION

Generative AI (GenAI) tools such as ChatGPT have spread rapidly through higher education. Students value them for explanations and writing support but also worry about accuracy, over-reliance and academic integrity (Chan and Hu, 2023; Cotton *et al.*, 2024; Schei *et al.*, 2024). Institutions deciding how to respond need to know how students see these tools, and researchers need sound instruments to measure those views.

Several instruments now exist. Some measure general attitudes towards AI (Grassini, 2023; Schepman and Rodway, 2020; Sindermann *et al.*, 2021) or competence in using it (Wang *et al.*, 2023a). Others measure students' acceptance or use of GenAI specifically, usually building on the Technology Acceptance Model or the Unified Theory of Acceptance and Use of Technology (UTAUT) (Davis, 1989; Venkatesh *et al.*, 2003), and several have been validated psychometrically (Kayalı *et al.*, 2026; Nemt-allah *et al.*, 2024; Ortega-Sánchez and Pérez-González, 2026). Chan and Zhou (2023) took a different route, basing an instrument on expectancy-value theory to capture the value and cost that students attach to GenAI.

Two gaps remain. First, acceptance models explain intentions mainly through beliefs about a tool's usefulness and ease of use, whereas theories of learning motivation emphasise how much students value a task and whether an activity meets their needs for autonomy, competence and relatedness (Eccles and Wigfield, 2020; Ryan and Deci, 2000). Few published instruments appear to bring these perspectives

together for GenAI. Second, validation studies of acceptance instruments typically stop at internal structure. They seldom test whether scores are stable over time or whether they predict behaviour, even though acceptance constructs are known to predict intentions more reliably than actual use (Bailey and Bista, 2026; Turner *et al.*, 2010).

This study reports the development and initial validation of the Values, Expectancy and Motivation Perception of AI in Education (VEM-AIEd) questionnaire, which integrates UTAUT, expectancy-value theory and self-determination theory and adds a scale on ethical considerations. In line with a unified view of validity that draws on several kinds of evidence (Messick, 1995), the study examines the questionnaire's internal structure, reliability, convergent and discriminant validity, stability over six weeks, and relations with self-reported AI use and logged use of a course chatbot. Because 18 of its 45 items were negatively worded, it also examines whether item wording shaped the results.

BACKGROUND

Theoretical Basis of the Questionnaire

UTAUT explains the intention to use a technology through four beliefs: that it will improve performance (performance expectancy), that it is easy to use (effort expectancy), that important others endorse its use (social influence), and that resources and support are available (facilitating conditions) (Venkatesh *et al.*, 2003). The model was later extended as UTAUT2 (Venkatesh *et al.*, 2012), and both versions have been applied to students' acceptance of ChatGPT (Strzelecki, 2024; Strzelecki and ElArabawy, 2024).

Expectancy-value theory holds that engagement depends on expecting to succeed and on valuing a task. Value can come from the task's importance to one's identity (attainment value), from enjoyment (intrinsic value) and from usefulness for future goals (utility value), and it is weighed against perceived cost (Eccles and Wigfield, 2020; Wigfield and Eccles, 2000). The theory has been used to measure students' perceptions of GenAI (Chan and Zhou, 2023) and to explain their intentions to learn about AI (Wang *et al.*, 2023b). Self-determination theory adds that motivation is sustained when activities satisfy the needs for autonomy, competence and relatedness (Ryan and Deci, 2000). Finally, because students' use of GenAI raises questions about integrity and responsible use (Chan and Hu, 2023; Cotton *et al.*, 2024), the questionnaire included a scale on ethical considerations.

Negatively Worded Items and Method Effects

Questionnaire designers often reverse the wording of some items to discourage patterned responding. The practice has costs. Negatively worded items can form a factor of their own even when they measure the same construct as positively worded items (DiStefano and Motl, 2006; Marsh, 1996), and a small share of careless respondents can be enough to produce such a factor (Schmitt and Stuits, 1985). Misresponse to reversed items has been traced to acquiescence, inattention and the extra effort needed to process negations (Swain *et al.*, 2008; Weijters *et al.*, 2013). Reviews have therefore questioned whether reversed items do more harm than good (Suárez-Álvarez *et al.*, 2018; van Sonderen *et al.*, 2013), and method variance of this kind can distort the relationships between constructs measured with the same instrument (Podsakoff *et al.*, 2003).

Separating a wording effect from substantive content requires each construct to be measured in more than one way, as in multitrait–multimethod designs (Eid, 2000). When every item of a scale shares one wording direction, wording and content are confounded, and a method factor cannot fully disentangle them. This was the situation in the present questionnaire.

METHOD

Questionnaire Development

The VEM-AIEd questionnaire was developed for a doctoral study of GenAI in sport science education (Ng, 2025). It has 45 items in 15 three-item scales rated from 1 (*strongly disagree*) to 5 (*strongly agree*). Five scales came from UTAUT (performance expectancy, effort expectancy, social influence, facilitating conditions and behavioural intention) and were adapted from Bani Ahmad *et al.* (2023). Six came from expectancy-value theory (confidence, success expectancy, attainment value, intrinsic value, utility value and cost) and were adapted from Chan and Zhou (2023). Three addressed the self-determination needs of autonomy, competence and relatedness (Ryan and Deci, 2000), and the ethical considerations scale was written from themes raised by lecturers and students in an earlier qualitative phase (Ng, 2025). The items were reviewed and refined with the doctoral supervisor and colleagues before use. To reduce patterned responding, all items in six scales were negatively worded:

Table 1: Sample Characteristics (N = 140).

Characteristic	*n*	%
Course level		
Certificate	30	21.4
Diploma	52	37.1
Degree	58	41.4
Gender		
Male	97	69.3
Female	43	30.7
Generation		
Gen X (born 1965–1980)	3	2.1
Gen Y (born 1981–1996)	16	11.4
Gen Z (born 1997–2012)	121	86.4
Survey term		
June 2024	92	65.7
September 2024	48	34.3
Prior AI use in studies		
Not at all	6	4.3
Rarely	30	21.4
A few times a month	23	16.4
Once a week	23	16.4
Several times a week	49	35.0
Daily	9	6.4

Note: Generation is based on the birth-year band each student selected. Prior AI use is self-reported frequency of using AI tools in their studies.

performance expectancy, effort expectancy, intrinsic value, facilitating conditions, behavioural intention and competence. Appendix A lists every item.

Participants and Procedure

Students in the certificate, diploma and degree programmes in sport and exercise science at a private higher education institution in Singapore were invited to complete the questionnaire online during class in the first week of the June and September 2024 terms. About 200 students were enrolled during these terms. Of 176 submissions, 3 were excluded because the student did not consent and 24 were later submissions from students who had already responded, leaving 149 students. A further 9 students gave the same answer to all 45 items and were removed as probable careless responders (Curran, 2016; Meade and Craig, 2012), giving an analytic sample of 140 with no missing data (Table 1).

Some of these students also took part in a six-week study in which they chose whether to use a course-specific AI chatbot (Ng, 2026). That study provided two further data sets used here: repeat questionnaires from 44 control-group students, for test–retest reliability, and the number of chatbot sessions logged for 41 chatbot users, for predictive validity. The study was approved by the Exeed College Ethics Committee in May 2024, and all participants gave informed consent.

Data Analysis

Item Analysis and Factorability

Item means, standard deviations, skewness, kurtosis and corrected item–total correlations were examined. Factorability was assessed with the Kaiser–Meyer–Olkin measure (Kaiser, 1974) and Bartlett's test of sphericity.

Exploratory Factor Analysis

Because the polychoric correlation matrix was not positive definite at this sample size, exploratory factor analysis (EFA) used Pearson correlations. Factors were extracted by minimum residuals with oblimin rotation, following recommended practice for common factor analysis (Fabrigar *et al.*, 1999). The number of factors was guided by parallel analysis (Horn, 1965) and the minimum average partial test (Velicer, 1976). Two-factor and 15-factor solutions were also inspected, to test the wording and the designed structures directly.

Confirmatory Factor Analysis

Mardia's tests indicated multivariate non-normality, so confirmatory factor analysis (CFA) used robust maximum likelihood (MLR) (Yuan and Bentler, 2000); with five response categories, robust continuous estimation performs comparably to categorical methods (Rhemtulla *et al.*, 2012). Four models were compared: one general factor (M1); two factors defined by wording direction (M2); the designed 15 correlated factors (M3); and M3 with an added method factor, uncorrelated with the substantive factors, for the 18 negatively worded items (M4). Hu and Bentler (1999) suggested CFI and TLI close to 0.95, $RMSEA$ close to 0.06 and $SRMR$ close to 0.08. Values of 0.90 for CFI and TLI and up to 0.08 for $RMSEA$ are widely treated as acceptable, and such cut-offs are guidelines rather than strict rules, especially for complex models (Marsh *et al.*, 2004). Nested models were compared with the scaled chi-square difference test (Satorra and Bentler, 2001). M3 was re-estimated including the straight-liners and with items treated as ordinal ($WLSMV$) as sensitivity checks.

Reliability and Validity

Internal consistency was estimated with Cronbach's α (Cronbach, 1951; Taber, 2018) and with composite reliability from the standardised CFA loadings, which does not assume equal loadings (McNeish, 2018). Convergent validity was judged by average variance extracted ($AVE \geq 0.50$), and discriminant validity by the Fornell–Larcker criterion (Fornell and Larcker, 1981) and the heterotrait–monotrait ratio ($HTMT < 0.85$) (Henseler *et al.*, 2015). Test–retest reliability over six weeks was estimated in the 44 control-group students with two-way random-effects, absolute-agreement, single-measure intraclass correlations (ICC(2,1) (Shrout and Fleiss, 1979), interpreted as poor (< 0.50), moderate (0.50–0.75), good (0.75–0.90) or excellent (> 0.90) (Koo and Li, 2016). Criterion validity was examined with Spearman correlations between scale scores and self-reported frequency of AI use in studies. Predictive validity was examined by correlating the 41 chatbot users' pre-test scale scores with the number of sessions the platform logged over the following six weeks.

Exploratory Revisions

Revised models were then estimated to guide the questionnaire's further development. Because they were specified after inspecting the data, they are hypotheses for testing in new samples rather than validated structures.

Sample Size and Software

A sample of 140 is modest for a 45-item, 15-factor model. Adequate sample size depends on communalities, the number of indicators per factor and model complexity (MacCallum *et al.*, 1999; Wolf *et al.*, 2013), and the sample was too small to split into separate exploratory and confirmatory halves, as is recommended in scale development (Boateng *et al.*, 2018; Worthington and Whittaker, 2006). Analyses used R 4.3.3 (R Core Team, 2024) with lavaan 0.6-17 (Rosseel, 2012), psych 2.4.1 (Revelle, 2024) and GPArotation 2024.2.1 (Bernaards and Jennrich, 2005).

RESULTS

Item Properties and Factorability

Item means ranged from 2.68 to 3.89 and standard deviations from 0.66 to 1.12 (Appendix A). Skewness (−1.11 to .07) and kurtosis (−0.65 to 1.78) were within ±2 for every item, but Mardia's tests showed multivariate non-normality ($skewness$ = 982.5, $kurtosis$ = 2365.4, both $p < 0.001$). Corrected item–total correlations ranged from 0.33 to 0.80; the lowest were for one cost item (0.33) and two social influence items (0.43 and 0.45). The Kaiser–Meyer–Olkin measure was 0.81, with item values from 0.60 to 0.91, and Bartlett's test was significant, $\chi^2(990) = 3980.9$, $p < 0.001$.

Table 2: Six-Factor Exploratory Solution: Items Loading ≥ .40.

Factor	Items (pattern loading)	Interpretation
F1	FC1 (0.59), FC2 (0.78), FC3 (0.71), BI1 (0.81), BI2 (0.74), BI3 (0.68), IV3 (0.55), COMP1 (0.59), COMP2 (0.66), COMP3 (0.65)	Negatively worded: facilitating conditions, intention, competence
F2	CON1 (0.42), CON2 (0.51), CON3 (0.41), COS2 (0.49), COS3 (0.52), EC1 (0.74), EC2 (0.67), EC3 (0.73)	Ethical considerations, confidence, cost
F3	CON1 (0.44), CON3 (0.43), SUC1 (0.40), SUC2 (0.45), AV1 (0.70), AV2 (0.86), AV3 (0.77), UV2 (0.45)	Attainment value, success, confidence
F4	PE1 (0.76), PE2 (0.66), PE3 (0.76), EE1 (0.61), EE2 (0.75), EE3 (0.66), IV2 (0.43)	Negatively worded: performance and effort expectancy
F5	SUC1 (0.54), SUC2 (0.41), COS1 (−0.42), AUT2 (0.45), AUT3 (0.40)	Success and autonomy (mixed)
F6	REL1 (0.70), REL2 (0.61), REL3 (0.62)	Relatedness

Note: Minimum residual extraction, oblimin rotation, Pearson correlations; 52% of variance explained. No loading ≥ .40: SI1, SI2, SI3, SUC3, IV1, UV1, UV3, AUT1. Item codes follow Appendix A; PE, EE, FC, BI, IV and COMP items are negatively worded.

Exploratory Factor Analysis

Parallel analysis suggested six factors and the minimum average partial test five, far fewer than the 15 designed scales. The two-factor solution divided the items by wording. All 18 negatively worded items loaded on one factor, and 21 of the 27 positively worded items loaded on the other (factor correlation = 0.25). The remaining positively worded items had no loading of at least 0.40, except one cost item, which loaded negatively (−0.40) on the negatively worded factor.

The six-factor solution explained 52% of the variance (Table 2). Two of its factors consisted entirely of negatively worded items: one combined facilitating conditions, behavioural intention and competence, and the other combined performance and effort expectancy, each with one intrinsic value item. The positively worded items formed factors for ethical considerations with confidence and two cost items, for attainment value with success expectancy and confidence, for a mix of success expectancy and autonomy items, and for relatedness. Eight items, including all three social influence items, had no loading of at least .40. Factor correlations ranged from −0.07 to 0.32.

When 15 factors were extracted (69% of variance), 12 of the designed scales formed a factor of their own, although only two of three items loaded at least 0.40 for facilitating conditions and for success expectancy. Performance and effort expectancy merged into a single factor, and the social influence items split across two factors. Five items cross-loaded at 0.30 or above.

Confirmatory Factor Analysis

The designed 15-factor model (M3) fitted far better than the one-factor and wording models (Table 3). Its $RMSEA$ and $SRMR$ were acceptable, but CFI and TLI fell below 0.90: $\chi^2(840) = 1263.1$, $CFI = 0.868$, $TLI = 0.845$, $RMSEA = 0.062$ [0.055, 0.069], $SRMR = 0.071$. Standardised loadings ranged from 0.40 to 0.88. Adding a negative-wording method factor (M4) did not improve fit significantly, $\Delta\chi^2(18) = 26.3$, $p = 0.092$, and produced an inadmissible solution; its loadings averaged 0.24 and ranged from −0.13 to 0.60. Results were similar when the straight-liners were included ($CFI = 0.879$, $RMSEA = 0.063$). Treating items as ordinal gave a higher CFI (0.897) but an inadmissible solution.

Reliability, Convergent and Discriminant Validity

Thirteen scales had α between 0.78 and 0.87, composite reliability between 0.79 and 0.88 and AVE between 0.56 and 0.70 (Table 4). Social influence ($\alpha = 0.66$, $composite\ reliability = 0.68$, $AVE = 0.42$) and cost ($\alpha = 0.61$, $composite\ reliability = 0.63$, $AVE = 0.37$) fell below conventional thresholds.

All $HTMT$ ratios were below 0.85 (Appendix B); the highest, 0.84, was between performance and effort expectancy. Their latent correlation (0.84) exceeded the square root of either scale's AVE (0.77 and 0.76), so the pair failed the Fornell–Larcker criterion, while every other scale met it. Merging the two scales nevertheless worsened fit significantly, $\Delta\chi^2(14) = 34.5$, $p = 0.002$, suggesting that they are closely related but not identical.

Item Wording and Latent Correlations

Latent correlations followed item wording more closely than content (Fig. 1). They averaged 0.53 among the six negatively worded scales and 0.35 among the nine positively worded scales, but only 0.11 between the two sets. Scales with closely related content but opposite wording were nearly

Table 3: Fit of Competing Confirmatory Factor Models.

Model	χ^2	df	CFI	TLI	$RMSEA$ [90% CI]	$SRMR$
M1. One general factor	3165.3	945	0.324	0.292	0.132 [0.127, 0.137]	0.151
M2. Two wording factors (positive, negative)	2415.1	944	0.525	0.502	0.110 [0.105, 0.116]	0.120
M3. 15 correlated factors (designed structure)	1263.1	840	0.868	0.845	0.062 [0.055, 0.069]	0.071
M4. M3 + negative-wording method factor[b]	1237.4	822	0.873	0.847	0.061 [0.054, 0.068]	0.072
Sensitivity analyses of M3						
M3 including straight-liners ($N = 149$)	1281.3	840	0.879	0.857	0.063	0.071
M3 with ordinal items ($WLSMV$)[b]	1446.2	840	0.897	0.879	0.072	0.077

Note: N = 140. Robust maximum likelihood (MLR) estimates with scaled χ^2 and robust CFI, TLI and $RMSEA$, except the $WLSMV$ model. All χ^2 $p < 0.001$. [b] Inadmissible solution (latent covariance matrix not positive definite).

Table 4: Scale statistics, reliability, convergent validity and six-week stability.

Scale	*M*	*SD*	*α*	*CR*	*AVE*	*λ* range	ICC(2,1) [95% *CI*]
UTAUT							
Performance expectancy[a]	3.63	0.91	0.81	0.81	0.58	0.72–0.83	0.23 [−0.07, 0.50]
Effort expectancy[a]	3.65	0.78	0.80	0.80	0.57	0.75–0.77	0.41 [0.13, 0.63]
Social influence	3.50	0.57	0.66	0.68	0.42	0.55–0.80	0.20 [−0.09, 0.46]
Facilitating conditions[a]	3.55	0.75	0.87	0.88	0.70	0.74–0.88	0.36 [0.09, 0.59]
Behavioural intention[a]	3.55	0.76	0.85	0.85	0.65	0.76–0.83	0.55 [0.31, 0.73]
Expectancy-value theory							
Confidence	3.58	0.75	0.85	0.86	0.67	0.71–0.88	0.31 [0.01, 0.55]
Success expectancy	3.53	0.68	0.83	0.83	0.62	0.70–0.87	0.20 [−0.10, 0.47]
Attainment value	3.15	0.80	0.85	0.86	0.67	0.77–0.88	0.36 [0.07, 0.60]
Intrinsic value[a]	3.53	0.74	0.78	0.79	0.56	0.68–0.85	0.38 [0.10, 0.61]
Utility value	3.72	0.64	0.85	0.85	0.66	0.76–0.84	0.24 [−0.05, 0.50]
Cost	3.12	0.74	0.61	0.63	0.37	0.40–0.77	0.47 [0.20, 0.67]
Self-determination theory							
Autonomy	3.85	0.59	0.81	0.82	0.61	0.67–0.87	0.38 [0.10, 0.60]
Competence[a]	3.46	0.75	0.83	0.83	0.62	0.75–0.82	0.40 [0.12, 0.62]
Relatedness	3.08	0.70	0.81	0.81	0.58	0.70–0.84	0.41 [0.14, 0.62]
Developed for this project							
Ethical considerations	3.80	0.66	0.83	0.83	0.63	0.69–0.85	0.53 [0.28, 0.71]

Note: $N = 140$ except ICC ($n = 44$ control-group students surveyed six weeks apart). Scale scores are item means (range 1–5). CR = composite reliability; AVE = average variance extracted; λ = standardised loadings in the 15-factor model (M3).
[a] Negatively worded items, reverse-scored.

unrelated: performance expectancy and utility value correlated 0.03, intrinsic value and attainment value −0.05, and facilitating conditions and social influence 0.13. Competence ("*I feel incapable of using Generative AI tools effectively*", reversed) correlated only 0.34 with confidence ("*I feel confident in my ability to use Generative AI tools effectively*").

Stability, Criterion and Predictive Validity

Over six weeks, ICC(2,1) ranged from 0.20 for social influence to 0.55 for behavioural intention (Table 4). Only

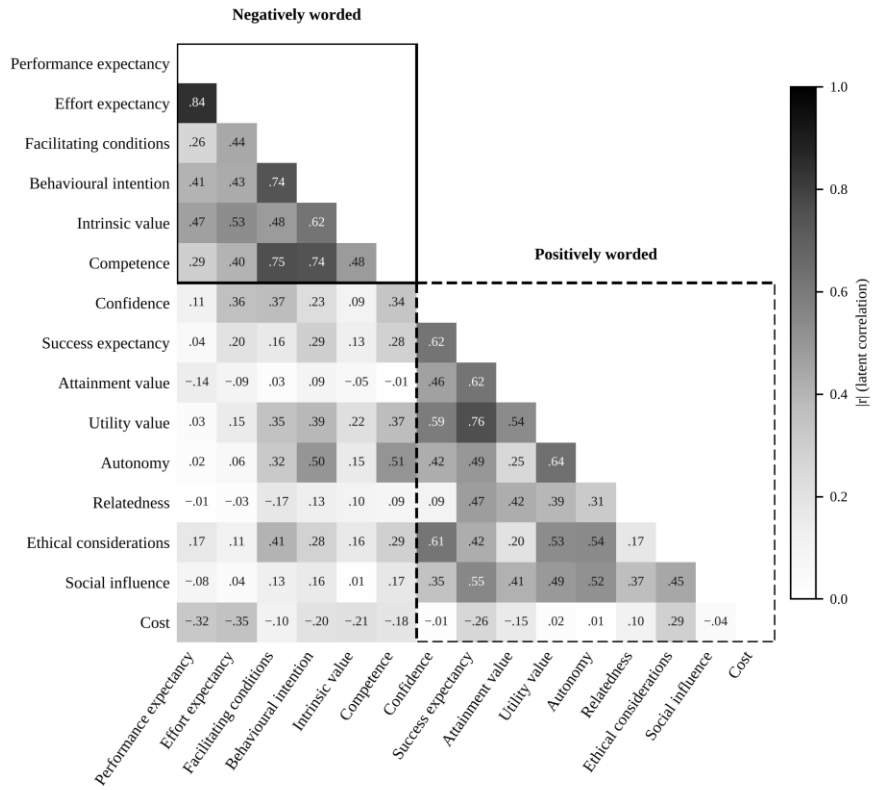

Fig. 1: Latent correlations between the 15 scales, grouped by item wording.

Table 5: Correlations of scale scores with self-reported AI use and logged Chatbot sessions.

Scale	ρ with AI use	p	ρ with sessions	p
Performance expectancy[a]	0.34	< 0.001	−0.15	0.361
Effort expectancy[a]	0.34	< 0.001	−0.01	0.957
Social influence	0.12	0.159	−0.15	0.358
Facilitating conditions[a]	0.21	0.015	−0.10	0.531
Behavioural intention[a]	0.34	<0 .001	−0.01	0.977
Confidence	0.47	< 0.001	−0.14	0.367
Success expectancy	0.30	< 0.001	−0.12	0.442
Attainment value	0.20	0.015	−0.08	0.618
Intrinsic value[a]	0.23	0.005	−0.28	0.076
Utility value	0.20	0.017	−0.11	0.477
Cost	−0.05	0.541	0.08	0.636
Autonomy	0.28	< 0.001	−0.07	0.672
Competence[a]	0.25	0.003	−0.01	0.949
Relatedness	0.11	0.189	−0.16	0.322
Ethical considerations	0.27	0.001	−0.22	0.169

Note: Spearman correlations. AI use: self-reported frequency of using AI tools in studies (N = 140). Sessions: chatbot sessions logged by the platform over six weeks, correlated with pre-test scores (n = 41). [a] Reverse-scored.

Table 6: Fit of exploratory revised models.

Model	χ^2	df	CFI	TLI	$RMSEA$ [90% CI]	$SRMR$
R1. Performance and effort expectancy merged (14 factors, 45 items)	1300.5	854	0.861	0.839	0.063 [0.056, 0.070]	0.073
R2. Social influence and cost removed (13 factors, 39 items)	883.8	624	0.909	0.892	0.057 [0.048, 0.065]	0.062
R3. R1 and R2 combined (12 factors, 39 items)	920.0	636	0.900	0.883	0.059 [0.050, 0.067]	0.065
R4. Positively worded scales only (9 factors, 27 items)	445.8	288	0.903	0.882	0.066 [0.054, 0.077]	0.071
R5. R4 without social influence and cost (7 factors, 21 items)	250.3	168	0.939	0.924	0.064 [0.047, 0.080]	0.058

Note: N = 140, MLR estimation. All solutions admissible. These models were specified after inspecting the data and require

behavioural intention and ethical considerations (0.53) reached the moderate range; the other scales were in the poor range.

Twelve scales correlated positively with self-reported frequency of AI use (ρ = 0.20–0.47, $p < 0.05$), most strongly confidence (ρ = 0.47) and behavioural intention, performance expectancy and effort expectancy (each ρ = 0.34) (Table 5). Social influence, relatedness and cost were unrelated to reported use. In contrast, no pre-test scale predicted the number of chatbot sessions logged over the following six weeks (all $|\rho| \leq 0.28$, all $p \geq 0.07$); for behavioural intention, $\rho = -0.01$.

Exploratory Revised Models

Merging performance and effort expectancy (R1) did not improve fit (Table 6). Removing social influence and cost (R2) brought fit into the acceptable range: CFI = 0.909, TLI = 0.892, $RMSEA$ = 0.057, $SRMR$ = 0.062. The best fit came from the seven positively worded scales without social influence and cost (R5; 21 items): $\chi^2(168) = 250.3$, CFI = 0.939, TLI = 0.924, $RMSEA$ = 0.064 [0.047, 0.080], $SRMR$ = 0.058. In this model, loadings ranged from 0.69 to 0.89, composite reliability from 0.81 to 0.86 and AVE from 0.58 to 0.67, and the highest latent correlation was 0.76.

DISCUSSION

This study evaluated a 45-item questionnaire that combines technology acceptance, expectancy-value beliefs, psychological needs and ethical considerations. The evidence is mixed. The designed scales are recognisable and most are internally consistent, but the questionnaire's structure is shaped as much by the direction of item wording as by content. Scores were also unstable over six weeks and did not predict how students actually used a course chatbot.

Instrument Coherence

The 15-factor model fitted far better than the alternatives; its loadings were substantial, and 13 of the 15 scales met conventional standards for internal consistency and convergent validity. All HTMT ratios fell below .85, and 12 of the 15 scales emerged when the designed number of factors was extracted. Most scales also related as expected to students' reported use of AI. Taken together, these results suggest that the items written for each construct largely cohere, a necessary first step for any new instrument (Boateng *et al.*, 2018).

The Wording Effect

The wording effect is the study's most important finding. Parallel analysis supported only six factors; a two-factor solution separated the items by wording; and scales that share content but differ in wording were nearly uncorrelated. Performance expectancy and utility value both concern the usefulness of GenAI, yet correlated 0.03. Competence and confidence contain almost identical statements, apart from negation, yet correlated only 0.34. This pattern matches earlier work showing that negatively worded items can form

factors of their own (DiStefano and Motl, 2006; Marsh, 1996), driven partly by careless or acquiescent responding and partly by the difficulty of processing negations (Schmitt and Stuits, 1985; Swain *et al.*, 2008; Weijters *et al.*, 2013).

The method-factor model could not resolve the problem. Because each of the six negatively worded scales was worded in one direction only, wording and content were confounded (Eid, 2000), and the method factor neither improved fit significantly nor produced an admissible solution. The practical consequence is that correlations between negatively and positively worded scales are likely to be underestimated, and correlations among negatively worded scales overestimated (Podsakoff *et al.*, 2003). Studies that use the full questionnaire to test relationships between constructs should therefore be interpreted with caution, and future versions should avoid mixing item directions across scales (Suárez-Álvarez *et al.*, 2018; van Sonderen *et al.*, 2013).

Weak Scales and Overlapping Constructs

Social influence and cost were the weakest scales, and both combine heterogeneous content. The social influence items refer to influential adults and teachers, institutional support and peers' use of GenAI, three sources that need not move together. The cost items combine mental effort, worries about critical thinking and concerns about privacy. Each scale could be rebuilt around a single, clearly defined facet. Performance and effort expectancy were highly correlated (0.84) but statistically separable; their overlap may reflect shared negative wording as much as shared content.

Stability and Prediction of Behaviour

Test–retest reliability was poor for most scales. Three explanations are possible and cannot be separated here: students' views of a fast-changing technology may genuinely shift within six weeks; wording effects add unstable variance; and the estimates rest on only 44 students. A shorter retest interval would help to distinguish measurement error from real change.

No scale predicted logged chatbot use, including behavioural intention. Several factors probably contributed. Use was low and skewed, which restricts the correlations that can be observed. The items asked about GenAI in general rather than the specific chatbot. And intentions are an imperfect guide to behaviour in general (Sheeran, 2002) and in technology acceptance in particular (Turner *et al.*, 2010). In a recent UTAUT2 study, the model explained 75% of the variance in students' intention to use ChatGPT but only 16.7% of the variance in use (Bailey and Bista, 2026), whereas intention strongly predicted reported use in a large Austrian sample (Kinskofer and Tulis, 2025). The contrast between the present results for reported and logged use underlines the value of validating acceptance instruments against recorded behaviour.

Recommendations for Further Development

Four changes would strengthen the questionnaire. First, all items should be worded in the same direction, or, if both directions are kept, each scale should contain items of both kinds so that wording effects can be modelled. Second, social influence and cost should be rewritten around single facets. Third, tool-specific versions should be developed when the aim is to predict use of a particular tool. Fourth, the revised questionnaire should be tested in larger samples split into exploratory and confirmatory halves, with tests of measurement invariance across groups (Boateng *et al.*, 2018; Worthington and Whittaker, 2006). The 21-item, seven-scale model, which fitted well and had strong loadings, offers a starting point. Because it was derived from the same data, however, it should not yet be described as validated.

LIMITATIONS

The sample came from one discipline at one private institution in Singapore and was mostly male and Generation Z, which limits generalisability. At 140 respondents, it was modest for a 15-factor model and too small to split for separate exploratory and confirmatory analyses, so both were run on the same data. The EFA used Pearson rather than polychoric correlations because the polychoric matrix was not positive definite. Test–retest and predictive analyses relied on small subsamples of 44 and 41 students, and chatbot use was low. Groups were too small to test measurement invariance by gender or course level. Finally, the revised models were specified after inspecting the data and are therefore exploratory.

CONCLUSION

The VEM-AIEd questionnaire measures a broad set of constructs relevant to students' acceptance of GenAI, and most of its scales are internally consistent. Its mixed item wording, however, substantially distorts the relationships between scales, two scales are weak, scores were unstable over six weeks, and none predicted actual use of a course chatbot. A shorter, consistently worded version shows promise but needs confirmation. More broadly, the study shows why developers of GenAI acceptance instruments should avoid mixing item directions and should validate scores against observed behaviour, not only against other self-reports.

ORCID iD

Ng Jun Wei : 0000-0001-8634-1938

Acknowledgement

The author thanks the students who took part in the study and Dr Mark Anthony Naval for supervising the doctoral research.

Grant Support Details

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Relationship to Other Work

This article is based on the author's Doctor of Education thesis (Ng, 2025). The test–retest and predictive analyses use data from a companion study of a course-specific chatbot (Ng, 2026), and both articles draw on the same pre-test responses. The analysis here was run on a cleaned data set, so some figures differ from those in the thesis.

Conflict of Interest

The author declares no competing interests. The course chatbot is anonymised and was used in the companion study and was provided free of charge by the company.

Ethics Approval

The study was approved in May 2024 by the Exeed College Ethics Committee in collaboration with Universidad Católica San Antonio de Murcia.

All participants gave informed consent online before completing the questionnaire. Responses submitted without consent were excluded.

Data Availability

The data are not publicly available because participants were told that their data would not be shared outside this research.

Use of Generative AI

The author used Claude (Anthropic), an AI assistant, to help with editing and rephrasing of text used in the manuscript. The author ran the analysis, edited all content, and takes full responsibility for the work.

REFERENCES

1) Bailey, K. and Bista, K. (2026) 'Habit before utility: understanding student acceptance and use of generative AI in higher education', *Journal of Interdisciplinary Studies in Education*, 15(7). https://doi.org/10.32674/9dyk2n48
2) Bani Ahmad, F., Al-Nawaiseh, S.J. and Al-Nawaiseh, A.J. (2023) 'Receptivity level of faculty members in universities using digital learning tools: a UTAUT perspective', *International Journal of Emerging Technologies in Learning*, 18(13), pp. 209-219. https://doi.org/10.3991/ijet.v18i13.39763
3) Bernaards, C.A. and Jennrich, R.I. (2005) 'Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis', *Educational and Psychological Measurement*, 65(5), pp. 676-696. https://doi.org/10.1177/0013164404272507
4) Boateng, G.O., Neilands, T.B., Frongillo, E.A., *et al.* (2018) 'Best practices for developing and validating scales for health, social, and behavioral research: a primer', *Frontiers in Public Health*, 6, 149. https://doi.org/10.3389/fpubh.2018.00149
5) Chan, C.K.Y. and Hu, W. (2023) 'Students' voices on generative AI: perceptions, benefits, and challenges in higher education', *International Journal of Educational Technology in Higher Education*, 20(1), 43. https://doi.org/10.1186/s41239-023-00411-8
6) Chan, C.K.Y. and Zhou, W. (2023) 'An expectancy value theory (EVT) based instrument for measuring student perceptions of generative AI', *Smart Learning Environments*, 10(1), 64. https://doi.org/10.1186/s40561-023-00284-4
7) Cotton, D.R.E., Cotton, P.A. and Shipway, J.R. (2024) 'Chatting and cheating: ensuring academic integrity in the era of ChatGPT', *Innovations in Education and Teaching International*, 61(2), pp. 228-239. https://doi.org/10.1080/14703297.2023.2190148
8) Cronbach, L.J. (1951) 'Coefficient alpha and the internal structure of tests', *Psychometrika*, 16(3), pp. 297-334. https://doi.org/10.1007/BF02310555
9) Curran, P.G. (2016) 'Methods for the detection of carelessly invalid responses in survey data', *Journal of Experimental Social Psychology*, 66, pp. 4-19. https://doi.org/10.1016/j.jesp.2015.07.006
10) Davis, F.D. (1989) 'Perceived usefulness, perceived ease of use, and user acceptance of information technology', *MIS Quarterly*, 13(3), pp. 319-340. https://doi.org/10.2307/249008
11) DiStefano, C. and Motl, R.W. (2006) 'Further investigating method effects associated with negatively worded items on self-report surveys', *Structural Equation Modeling*, 13(3), pp. 440-464. https://doi.org/10.1207/s15328007sem1303_6
12) Eccles, J.S. and Wigfield, A. (2020) 'From expectancy-value theory to situated expectancy-value theory: a developmental, social cognitive, and sociocultural perspective on motivation', *Contemporary Educational Psychology*, 61, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859
13) Eid, M. (2000) 'A multitrait-multimethod model with minimal assumptions', *Psychometrika*, 65(2), pp. 241-261. https://doi.org/10.1007/BF02294377
14) Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., *et al.* (1999) 'Evaluating the use of exploratory factor analysis in psychological research', *Psychological Methods*, 4(3), pp. 272-299. https://doi.org/10.1037/1082-989X.4.3.272
15) Fornell, C. and Larcker, D.F. (1981) 'Evaluating structural equation models with unobservable variables and measurement error', *Journal of Marketing Research*, 18(1), pp. 39-50. https://doi.org/10.2307/3151312
16) Grassini, S. (2023) 'Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence', *Frontiers in Psychology*, 14, 1191628. https://doi.org/10.3389/fpsyg.2023.1191628
17) Henseler, J., Ringle, C.M. and Sarstedt, M. (2015) 'A new criterion for assessing discriminant validity in variance-based structural equation modeling', *Journal of the Academy of Marketing Science*, 43(1), pp. 115-135. https://doi.org/10.1007/s11747-014-0403-8
18) Horn, J.L. (1965) 'A rationale and test for the number of factors in factor analysis', *Psychometrika*, 30(2), pp. 179-185. https://doi.org/10.1007/BF02289447
19) Hu, L. and Bentler, P.M. (1999) 'Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives', *Structural Equation Modeling*, 6(1), pp. 1-55. https://doi.org/10.1080/10705519909540118
20) Kaiser, H.F. (1974) 'An index of factorial simplicity', *Psychometrika*, 39(1), pp. 31-36. https://doi.org/10.1007/BF02291575
21) Kayalı, B., Yavuz, M., Üstün, A.G., *et al.* (2026) 'Development and validation of AI acceptance scale (AI-AS) for learners in higher education', *TechTrends*, 70(4), pp. 983-997. https://doi.org/10.1007/s11528-026-01209-x
22) Kinskofer, F. and Tulis, M. (2025) 'Motivational and appraisal factors shaping generative AI use and intention in Austrian higher education students and teachers', *Frontiers in Education*, 10, 1677827. https://doi.org/10.3389/feduc.2025.1677827
23) Koo, T.K. and Li, M.Y. (2016) 'A guideline of selecting and reporting intraclass correlation coefficients for reliability research', *Journal of Chiropractic Medicine*, 15(2), pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012
24) MacCallum, R.C., Widaman, K.F., Zhang, S., *et al.* (1999) 'Sample size in factor analysis', *Psychological Methods*, 4(1), pp. 84-99. https://doi.org/10.1037/1082-989X.4.1.84
25) Marsh, H.W. (1996) 'Positive and negative global self-esteem: a substantively meaningful distinction or artifactors?', *Journal of Personality and Social Psychology*, 70(4), pp. 810-819. https://doi.org/10.1037/0022-3514.70.4.810
26) Marsh, H.W., Hau, K.-T. and Wen, Z. (2004) 'In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings', *Structural Equation Modeling*, 11(3), pp. 320-341. https://doi.org/10.1207/s15328007sem1103_2
27) McNeish, D. (2018) 'Thanks coefficient alpha, we'll take it from here', *Psychological Methods*, 23(3), pp. 412-433. https://doi.org/10.1037/met0000144
28) Meade, A.W. and Craig, S.B. (2012) 'Identifying careless responses in survey data', *Psychological Methods*, 17(3), pp. 437-455. https://doi.org/10.1037/a0028085
29) Messick, S. (1995) 'Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning', *American Psychologist*, 50(9), pp. 741-749. https://doi.org/10.1037/0003-066X.50.9.741
30) Nemt-allah, M., Khalifa, W., Badawy, M., *et al.* (2024) 'Validating the ChatGPT usage scale: psychometric properties and factor structures among postgraduate students', *BMC Psychology*, 12(1), 497. https://doi.org/10.1186/s40359-024-01983-4
31) Ng, J.W. (2025) *Technological acceptance and motivation of generative artificial intelligence in sport science education at private higher education*. Doctoral thesis, Exeed College and Universidad Católica San Antonio de Murcia, Murcia, Spain.
32) Ng, J.W. (2026) 'Available but seldom used: a quasi-experiment of a course-specific AI chatbot in sport science education'. Manuscript in preparation, PSB Academy, Singapore.

33) Ortega-Sánchez, D. and Pérez-González, C. (2026) 'Psychometric validation of the 'ChatGPT acceptance and use scale' (AUS-GPT), and comparative analysis of its adoption among Mexican higher education students (2023-2025)', *Social Sciences & Humanities Open*, 13, 102717. https://doi.org/10.1016/j.ssaho.2026.102717
34) Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., *et al.* (2003) 'Common method biases in behavioral research: a critical review of the literature and recommended remedies', *Journal of Applied Psychology*, 88(5), pp. 879-903. https://doi.org/10.1037/0021-9010.88.5.879
35) R Core Team (2024) *R: a language and environment for statistical computing*. Version 4.3.3. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/
36) Revelle, W. (2024) *Psych: procedures for psychological, psychometric, and personality research*. R package version 2.4.1. Evanston, Illinois, USA: Northwestern University. https://CRAN.R-project.org/package=psych
37) Rhemtulla, M., Brosseau-Liard, P.É. and Savalei, V. (2012) 'When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions', *Psychological Methods*, 17(3), pp. 354-373. https://doi.org/10.1037/a0029315
38) Rosseel, Y. (2012) 'lavaan: an R package for structural equation modeling', *Journal of Statistical Software*, 48(2), pp. 1-36. https://doi.org/10.18637/jss.v048.i02
39) Ryan, R.M. and Deci, E.L. (2000) 'Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being', *American Psychologist*, 55(1), pp. 68-78. https://doi.org/10.1037/0003-066X.55.1.68
40) Satorra, A. and Bentler, P.M. (2001) 'A scaled difference chi-square test statistic for moment structure analysis', *Psychometrika*, 66(4), pp. 507-514. https://doi.org/10.1007/BF02296192
41) Schei, O.M., Møgelvang, A. and Ludvigsen, K. (2024) 'Perceptions and use of AI chatbots among students in higher education: a scoping review of empirical studies', *Education Sciences*, 14(8), 922. https://doi.org/10.3390/educsci14080922
42) Schepman, A. and Rodway, P. (2020) 'Initial validation of the general attitudes towards artificial intelligence scale', *Computers in Human Behavior Reports*, 1, 100014. https://doi.org/10.1016/j.chbr.2020.100014
43) Schmitt, N. and Stuits, D.M. (1985) 'Factors defined by negatively keyed items: the result of careless respondents?', *Applied Psychological Measurement*, 9(4), pp. 367-373. https://doi.org/10.1177/014662168500900405
44) Sheeran, P. (2002) 'Intention-behavior relations: a conceptual and empirical review', *European Review of Social Psychology*, 12(1), pp. 1-36. https://doi.org/10.1080/14792772143000003
45) Shrout, P.E. and Fleiss, J.L. (1979) 'Intraclass correlations: uses in assessing rater reliability', *Psychological Bulletin*, 86(2), pp. 420-428. https://doi.org/10.1037/0033-2909.86.2.420
46) Sindermann, C., Sha, P., Zhou, M., *et al.* (2021) 'Assessing the attitude towards artificial intelligence: introduction of a short measure in German, Chinese, and English language', *KI – Künstliche Intelligenz*, 35(1), pp. 109-118. https://doi.org/10.1007/s13218-020-00689-0
47) Strzelecki, A. (2024) 'Students' acceptance of ChatGPT in higher education: an extended unified theory of acceptance and use of technology', *Innovative Higher Education*, 49(2), pp. 223-245. https://doi.org/10.1007/s10755-023-09686-1
48) Strzelecki, A. and ElArabawy, S. (2024) 'Investigation of the moderation effect of gender and study level on the acceptance and use of generative AI by higher education students: comparative evidence from Poland and Egypt', *British Journal of Educational Technology*, 55(3), pp. 1209-1230. https://doi.org/10.1111/bjet.13425
49) Suárez-Álvarez, J., Pedrosa, I., Lozano, L., *et al.* (2018) 'Using reversed items in Likert scales: a questionable practice', *Psicothema*, 30(2), pp. 149-158. https://doi.org/10.7334/psicothema2018.33
50) Swain, S.D., Weathers, D. and Niedrich, R.W. (2008) 'Assessing three sources of misresponse to reversed Likert items', *Journal of Marketing Research*, 45(1), pp. 116-131. https://doi.org/10.1509/jmkr.45.1.116
51) Taber, K.S. (2018) 'The use of Cronbach's alpha when developing and reporting research instruments in science education', *Research in Science Education*, 48(6), pp. 1273-1296. https://doi.org/10.1007/s11165-016-9602-2
52) Turner, M., Kitchenham, B., Brereton, P., *et al.* (2010) 'Does the technology acceptance model predict actual use? A systematic literature review', *Information and Software Technology*, 52(5), pp. 463-479. https://doi.org/10.1016/j.infsof.2009.11.005
53) van Sonderen, E., Sanderman, R. and Coyne, J.C. (2013) 'Ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain', *PLoS ONE*, 8(7), e68967. https://doi.org/10.1371/journal.pone.0068967
54) Velicer, W.F. (1976) 'Determining the number of components from the matrix of partial correlations', *Psychometrika*, 41(3), pp. 321-327. https://doi.org/10.1007/BF02293557
55) Venkatesh, V., Morris, M.G., Davis, G.B., *et al.* (2003) 'User acceptance of information technology: toward a unified view', *MIS Quarterly*, 27(3), pp. 425-478. https://doi.org/10.2307/30036540
56) Venkatesh, V., Thong, J.Y.L. 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), pp. 157-178. https://doi.org/10.2307/41410412
57) Wang, B., Rau, P.-L.P. and Yuan, T. (2023a) 'Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale', *Behaviour & Information Technology*, 42(9), pp. 1324-1337. https://doi.org/10.1080/0144929X.2022.2072768
58) Wang, F., King, R.B., Chai, C.S., *et al.* (2023b) 'University students' intentions to learn artificial intelligence: the roles of supportive environments and expectancy-value beliefs', *International Journal of Educational Technology in Higher Education*, 20(1), 51. https://doi.org/10.1186/s41239-023-00417-2
59) Weijters, B., Baumgartner, H. and Schillewaert, N. (2013) 'Reversed item bias: an integrative model', *Psychological Methods*, 18(3), pp. 320-334. https://doi.org/10.1037/a0032121
60) Wigfield, A. and Eccles, J.S. (2000) 'Expectancy-value theory of achievement motivation', *Contemporary Educational Psychology*, 25(1), pp. 68-81. https://doi.org/10.1006/ceps.1999.1015
61) Wolf, E.J., Harrington, K.M., Clark, S.L., *et al.* (2013) 'Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety', *Educational and Psychological Measurement*, 73(6), pp. 913-934. https://doi.org/10.1177/0013164413495237
62) Worthington, R.L. and Whittaker, T.A. (2006) 'Scale development research: a content analysis and recommendations for best practices', *The Counseling Psychologist*, 34(6), pp. 806-838. https://doi.org/10.1177/0011000006288127
63) Yuan, K.-H. and Bentler, P.M. (2000) 'Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data', *Sociological Methodology*, 30(1), pp. 165-200. https://doi.org/10.1111/0081-1750.00078