

Development and Psychometric Validation of the Non-Academic Qualities (NAQ) Scale and Companion Vignette Battery for MBBS Graduates: A Multi-Phase Instrument Validation Study

Anbarasi Muthusamy¹, Rajkumar Gajendran^{2*}

¹ Professor of Physiology, Chettinad Hospital & Research Institute (Chettinad Academy of Research and Education), Chettinad Health City, SH 49A, Kelambakkam, Tamil Nadu 603103, India.

² Professor of Paediatrics, Chettinad Hospital & Research Institute (Chettinad Academy of Research and Education), Chettinad Health City, SH 49A, Kelambakkam, Tamil Nadu 603103, India.

*Corresponding author: Rajkumar Gajendran, grkmedico@gmail.com,

HOW TO CITE:

Anbarasi Muthusamy, Rajkumar Gajendran (2026). Development and Psychometric Validation of the Non-Academic Qualities (NAQ) Scale and Companion Vignette Battery for MBBS Graduates: A Multi-Phase Instrument Validation Study. International Journal of Special Education, 41(14s), 339-352

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

Background: Non-academic qualities such as professionalism, communication, ethical reasoning, resilience, and self-management are increasingly recognised as essential determinants of physician competence. However, no existing instrument combines a structured, multi-domain importance scale with vignette-based behavioural validation.

Objectives: This study aimed to develop and psychometrically validate the 24-item Non-Academic Qualities (NAQ) scale and its companion five-vignette battery for assessing non-academic qualities of MBBS graduates.

Methods: A six-phase design was employed: expert content validity assessment (n = 10), pilot testing of clarity and relevance (n = 50), internal consistency analysis, exploratory factor analysis (EFA), test-retest reliability, and expert inter-rater agreement on the vignette scoring key.

Key Results: All 24 items achieved an Item-Content Validity Index (I-CVI) ≥ 0.80 (Scale-Content Validity Index, S-CVI = 0.82). Pilot testing confirmed excellent face validity (mean Clarity Index = 0.958; mean Relevance Index = 0.936). Overall Cronbach's alpha was 0.939, indicating strong internal consistency. EFA yielded a five-factor solution explaining 66.9% of total variance, closely mirroring the a priori domain structure. Test-retest reliability was good (overall ICC = 0.834), and expert agreement on the vignette scoring key was almost perfect (Fleiss' kappa = 0.823).

Conclusion: The NAQ scale and its vignette battery demonstrate strong, convergent evidence of validity and reliability, supporting their use as a methodologically sound instrument for assessing non-academic qualities of MBBS graduates.

Keywords: Non-academic qualities, medical professionalism, instrument validation, content validity, factor analysis.

1. Introduction

Medical competence has traditionally been defined and assessed through cognitive and clinical knowledge benchmarks, yet a substantial body of evidence indicates that academic excellence alone is insufficient to produce a competent, trustworthy physician (Albanese et al., 2003; Patterson et al., 2016). Non-academic qualities, encompassing professionalism, communication, empathy, ethical reasoning, resilience and self-management, have increasingly been recognised as core determinants of physician performance, patient safety and the overall quality of healthcare delivery (General Medical Council, 2013; Frank et al., 2015).

Regulatory and accreditation bodies, including the General Medical Council's Good Medical Practice framework and the CanMEDS physician competency framework, now explicitly embed non-academic domains, such as professionalism, collaboration and communication, alongside medical expertise as essential physician roles (Frank et al., 2015; General Medical Council, 2013).

Despite this consensus on their importance, the literature reveals persistent ambiguity regarding which non-academic qualities matter most, and from whose perspective. Lambe and Bristow (2010) conducted a Delphi survey of ten clinicians to identify the most important non-academic attributes of good doctors, generating a ranked list evaluated on a five-point Likert scale. However, this study captured the views of clinicians alone, leaving the perspectives of trainees, allied health professionals and patients largely unexamined. Subsequent multi-stakeholder enquiries have begun to address this gap. Comparative surveys of medical students, doctors, nurses and patients have shown that stakeholder groups do not always converge on what constitutes the single most important physician attribute, and these discrepancies can have direct implications for patient-centred care delivery. Likewise, studies comparing faculty and student perceptions of educator qualities have demonstrated systematic divergence in priorities

between teachers and learners, underscoring that perception of what matters is contingent on stakeholder vantage point rather than fixed or self-evident.

A second limitation of the existing literature is methodological. Most published instruments assessing non-academic physician qualities rely exclusively on self-report Likert ratings of perceived importance, without a parallel assessment of applied decision-making under realistic conditions.

Vignette-based and situational judgement methodologies have been developed and separately validated as reliable, content-valid tools for assessing non-cognitive attributes in medical admissions, with reported inter-rater reliabilities as high as kappa = 0.96 and generalizability coefficients up to 0.81. Yet these have typically been deployed as stand-alone selection instruments rather than as a complementary validation layer alongside a structured importance-rating survey.

The integration of a stated-preference measure (Likert importance ratings) with a revealed-preference measure (vignette-based applied judgement) offers a methodological advantage. It permits triangulation between what respondents say they value and how they would actually act, a distinction with established relevance in attitude-behaviour research and one rarely operationalised in the non-academic competency literature to date.

A third gap concerns breadth of stakeholder representation. No identified instrument has incorporated the full ecosystem of individuals who interact with, train, supervise or are cared for by MBBS graduates within a single validated framework, spanning medical students across multiple training stages, postgraduate residents, medical faculty, staff nurses, allied health students and members of the public.

This breadth matters: non-academic qualities are, by definition, relational and contextual, and a physician's communication, empathy or accountability is enacted differently, and evaluated differently, depending on whether the observer is a supervising consultant, a co-

working nurse or a parent at a child's bedside. A multi-stakeholder framework is therefore not merely additive but conceptually necessary to capture the full construct.

The present study was designed to address these three gaps by developing and psychometrically validating the Non-Academic Qualities (NAQ) scale, a 24-item, five-domain instrument assessing the perceived importance of non-academic qualities in MBBS graduates, accompanied by a five-vignette behavioural validation battery. The objectives of this validation phase were: to establish the content validity of the 24-item scale through expert panel review; to confirm face validity and item comprehensibility through pilot testing in the target population; to evaluate internal consistency and underlying factor structure through Cronbach's alpha and exploratory factor analysis; to establish temporal stability through test-retest reliability; and to validate the companion vignette battery through expert inter-rater agreement on a defensible scoring key. Establishing this psychometric foundation is a prerequisite for the subsequent main-phase administration of the NAQ scale across twelve stakeholder groups. It also represents, to our knowledge, the first attempt to combine a structured multi-domain importance scale with vignette-based behavioural validation within a single instrument for the assessment of non-academic qualities in medical education.

2. Methodology

Design

This study employed a sequential, multi-phase instrument validation design comprising six stages: expert panel content validity assessment, pilot testing of clarity and respondent-perceived relevance, internal consistency analysis, exploratory factor analysis, test-retest reliability assessment, and expert inter-rater agreement on the vignette scoring key. This staged approach follows established recommendations for the development of psychometric instruments in health professions education (Artino et al., 2014; Polit & Beck, 2006).

Participants/Sample

Three distinct samples were recruited across the six phases:

- **Expert panel (n = 10):** Occupational therapists, clinical psychologists, senior faculty from recognised medical colleges, and practising clinicians with expertise in medical education and professionalism. This panel conducted both the Phase I content validity review and the Phase VI vignette scoring key validation.
- **Pilot sample (n = 50):** Respondents drawn from the intended target population, recruited to assess item comprehensibility and perceived relevance from a lay respondent perspective (Phase II).
- **Test-retest sample:** The same respondent sample completed the NAQ importance scale on two occasions, separated by a two-week interval (Phase V).

Instruments

- **NAQ scale:** A 24-item scale developed through review of existing literature on physician professionalism and non-academic competency frameworks, including the General Medical Council's Good Medical Practice guidance and the CanMEDS physician competency framework, supplemented by items from prior Delphi and survey-based studies of physician attributes (Lambe & Bristow, 2010). Items were organised into five conceptually derived domains: Domain 1, core professionalism and ethics (six items); Domain 2, communication and interpersonal skills (four items); Domain 3, personal and professional development (six items); Domain 4, professional adaptability and reflective practice (four items); Domain 5, modern competencies and self-management (four items). Full item wording is in Supplementary Table S1.
- **Vignette battery:** A complementary five-scenario vignette battery (V1–V5), each with two multiple-choice questions (Q1, Q2) and five response options, assessing applied professionalism judgement across ethical decision-making, self-care, cultural sensitivity, digital literacy and communication.

Procedure

Phase I — Content validity (expert panel):

Each expert independently rated all 24 items on a four-point relevance scale (1 = not relevant to 4 = highly relevant). The Item-level Content Validity Index (I-CVI) was the proportion of experts rating an item 3 or 4 (Lynn, 1986); the Scale-level Content Validity Index (S-CVI) was the average of all I-CVI values (Polit & Beck, 2006). Thresholds of $I-CVI \geq 0.78$ and $S-CVI \geq 0.80$ were set a priori.

Phase II — Pilot testing (clarity and relevance):

The 24-item scale was administered to the pilot sample. Each respondent classified every item as clear, somewhat clear, or confusing; the Clarity Index (CI) was the proportion selecting "clear" (threshold $CI \geq 0.80$). Respondents also rated relevance on the same four-point scale, yielding an Item Relevance Index ($RI \geq 0.78$ threshold).

Phase III — Internal consistency: Cronbach's coefficient alpha (Cronbach, 1951) was calculated overall and by domain. Corrected item-total correlations and alpha-if-item-deleted values identified weak-contributing items.

Phase IV — Construct validity (EFA):

Exploratory factor analysis with maximum likelihood extraction and varimax rotation. Sampling adequacy was checked via the Kaiser-Meyer-Olkin (KMO) measure (≥ 0.60 threshold) and Bartlett's test of sphericity ($p < 0.05$). Factors were extracted using the Kaiser criterion (eigenvalue > 1); loadings below 0.40 were suppressed. The empirical structure was compared against the a priori five-domain model.

Phase V — Test-retest reliability: The NAQ scale was re-administered after a two-week interval. The Intraclass Correlation Coefficient (ICC), two-way mixed-effects model, absolute agreement, single measures — ICC(2,1) (Shrout & Fleiss, 1979) — was calculated by domain and overall.

Phase VI — Vignette validation (expert inter-rater agreement): The same ten-member expert panel independently scored the five

vignettes (ten questions total), selecting the most appropriate response per question. For V3.Q1, where two options were pre-specified as equally acceptable, these were collapsed into one category before recalculating item-level agreement.

Data Analysis

All analyses used standard statistical computing methods. Descriptive statistics (means, SDs) were calculated for all items. Given the validation/pilot nature of this phase, results are reported descriptively without formal hypothesis testing; pre-specified thresholds guided item retention and instrument finalisation decisions.

- **Content validity:** $I-CVI \geq 0.78$, $S-CVI \geq 0.80$.
- **Internal consistency:** Cronbach's alpha ≥ 0.70 minimum, ≥ 0.90 excellent (George & Mallery, 2003; Nunnally & Bernstein, 1994).
- **Construct validity:** $KMO \geq 0.60$ (Kaiser, 1974), Bartlett's $p < 0.05$.
- **Test-retest reliability:** ICC interpreted per Koo & Li (2016) — < 0.50 poor, $0.50-0.75$ moderate, $0.75-0.90$ good, > 0.90 excellent; ICC ≥ 0.70 minimum acceptable.
- **Inter-rater agreement:** Fleiss' kappa (Fleiss, 1971); kappa > 0.61 required for substantial agreement (Landis & Koch, 1977).

3. Results

3.1 Expert Panel Composition

Content validation was conducted by a panel of ten domain experts comprising occupational therapists, clinical psychologists, senior faculty members from recognised medical colleges, and practising clinicians with expertise in medical education and professionalism. All ten experts independently evaluated each of the 24 items of the Non-Academic Qualities (NAQ) scale for relevance, clarity and domain representativeness.

3.2 Item-Level Content Validity, Clarity, Relevance and Reliability

Table 1 presents the consolidated item-level psychometric profile of the 24-item NAQ scale, combining the Phase I Item-level Content Validity Index (I-CVI), Phase II pilot Clarity Index (CI) and Relevance Index (RI), and Phase III descriptive statistics and reliability indices. Full item wording is provided in Supplementary Table S1.

Content validity: All 24 items achieved I-CVI values ≥ 0.80 . Four items (Items 1, 6, 17, 18, 24) achieved I-CVI = 0.90; the remaining 20 items achieved I-CVI = 0.80. The Scale-level Content Validity Index (S-CVI) was 0.82.

Clarity and relevance: All 24 items met or exceeded CI ≥ 0.80 . Twelve items achieved CI = 1.00. Four items (Item 1, Item 6, Item 21, Item 24) recorded the lowest CI values (each CI = 0.87). Mean CI across all items was 0.958. Item Relevance Indices ranged from RI = 0.93 (23 items) to RI = 1.00 (Item 9); mean RI was 0.936. No items fell below the RI ≥ 0.78 threshold.

Internal consistency: Overall Cronbach's alpha for the 24-item scale was 0.939. Corrected item-total correlations ranged from 0.364 to 0.730 for 23 of the 24 items. Item 9 (teamwork) recorded a corrected item-total correlation of 0.159; removal of Item 9 would increase overall alpha to 0.942. Domain-level alpha values were: Domain 1 = 0.946; Domain 2 = 0.855; Domain 3 = 0.943; Domain 4 = 0.916; Domain 5 = 0.926.

3.3 Construct Validity: Exploratory Factor Analysis

The Kaiser-Meyer-Olkin measure of sampling adequacy was KMO = 0.847. Bartlett's test of

sphericity was statistically significant ($\chi^2(276) = 1018.01$, $p < .001$). Using maximum likelihood extraction with varimax rotation, five factors were identified (Kaiser criterion, eigenvalue > 1), collectively explaining 66.9% of total variance.

- Factor 1 — Core professionalism and ethics (Items 1–6): loadings 0.692–0.879
- Factor 2 — Personal and professional development (Items 11–16): loadings 0.582–0.824
- Factor 3 — Interpersonal and communication skills (Items 7–10): loadings 0.499–0.720
- Factor 4 — Professional competencies and self-management (Items 20–24): loadings 0.617–0.886, h^2 up to 0.942
- Factor 5 — Reflective practice and lifelong learning (Items 17–19): loadings 0.676–0.765

Four of the five a priori domains showed exact correspondence with the extracted factors. Item 20 (time management and punctuality) showed a cross-loading tendency toward Factor 4. As a result, the item ranges in Table 1 and Table 2 do not align exactly for this domain: Table 1's a priori Domain 4 (professional adaptability and reflective practice) groups Items 17–20, whereas Table 2's rotated factor structure places Item 20 with Items 21–24 under Factor 4 (professional competencies and self-management), leaving Items 17–19 alone under Factor 5 (reflective practice and lifelong learning); Item 20 was retained within its original domain on theoretical grounds, as discussed in Section 4.1.

Table 1. Consolidated item-level psychometric profile of the NAQ scale (24 items)

Item	Description	I-CVI	CI	RI	Mean	SD	Corrected r	α if deleted
Domain 1 — Core professionalism and ethics								
1	Patient care priority	0.9	0.87	0.93	3.56	0.93	0.686	0.935
2	Honesty/integrity	0.8	0.93	0.93	3.6	1.07	0.73	0.934
3	Ethical decision-making	0.8	0.93	0.93	3.96	1.05	0.613	0.936

Item	Description	I-CVI	CI	RI	Mean	SD	Corrected r	α if deleted
4	Accountability for errors	0.8	0.93	0.93	3.86	0.99	0.579	0.937
5	Privacy/confidentiality	0.8	0.93	0.93	4.08	0.9	0.602	0.936
6	Professional dress code	0.9	0.87	0.93	3.76	1.02	0.67	0.935
Domain 2 — Communication and interpersonal skills								
7	Communication/listening skills	0.8	1	0.93	3.62	0.99	0.364	0.94
8	Empathy, non-judgemental attitude	0.8	1	0.93	3.86	0.76	0.562	0.937
9	Teamwork	0.8	1	1	3.54	0.84	0.159*	0.942
10	Cultural/religious respect	0.8	1	0.93	4.04	0.75	0.454	0.938
Domain 3 — Personal and professional development								
11	Motivation/commitment	0.8	0.93	0.93	4	0.97	0.624	0.936
12	Resilience	0.8	0.93	0.93	3.8	0.97	0.667	0.935
13	Flexibility/adaptability	0.8	1	0.93	3.94	0.77	0.604	0.936
14	Creativity/innovative thinking	0.8	1	0.93	3.98	0.84	0.569	0.937
15	Curiosity/inquisitiveness	0.8	1	0.93	3.44	0.99	0.636	0.936
16	Positive professional attitude	0.8	1	0.93	4.08	0.9	0.647	0.936
Domain 4 — Professional adaptability and reflective practice								
17	Recognition of own limits	0.9	1	0.93	3.8	0.99	0.719	0.935
18	Coping with ambiguity	0.9	1	0.93	3.92	0.99	0.669	0.935
19	Lifelong learning commitment	0.8	1	0.93	3.66	0.98	0.666	0.935
20	Time management/punctuality	0.8	1	0.93	4.08	0.85	0.635	0.936
Domain 5 — Modern competencies and self-management								
21	Leadership quality	0.8	0.87	0.93	3.52	1.07	0.724	0.934
22	Body language management	0.8	1	0.93	3.38	1.09	0.685	0.935
23	Self-care/wellbeing	0.8	0.93	0.93	3.96	0.83	0.585	0.937
24	Digital/health literacy	0.9	0.87	0.93	3.48	1.09	0.624	0.936
	Overall scale	0.82	0.958	0.936	—	—	—	0.939

Note. I-CVI = Item-level Content Validity Index (expert panel, $n = 10$); CI = Clarity Index; RI = Relevance Index (pilot sample, $n = 50$). Overall scale I-CVI value shown is the Scale-level Content Validity Index (S-CVI = 0.82). *Item 9 corrected item-total $r = 0.159$, below the conventional 0.20 threshold. Full item wording is provided in Supplementary Table S1.

Table 2. Varimax-rotated factor loadings and communalities (maximum likelihood extraction)

Item	Description	F1	F2	F3	F4	F5	h ²
Factor 1 — Core professionalism and ethics (Items 1–6)							
1	Patient care priority	0.754	—	—	—	—	0.69
2	Honesty/integrity	0.851	—	—	—	—	0.893
3	Ethical decision-making	0.879	—	—	—	—	0.833
4	Accountability for errors	0.821	—	—	—	—	0.713
5	Privacy/confidentiality	0.692	—	—	—	—	0.542
6	Professional dress code	0.794	—	—	—	—	0.779
Factor 2 — Personal and professional development (Items 11–16)							
11	Motivation/commitment	—	0.798	—	—	—	0.721
12	Resilience	—	0.824	—	—	—	0.812
13	Flexibility/adaptability	—	0.582	—	—	—	0.394
14	Creativity/innovative thinking	—	0.652	—	—	—	0.477
15	Curiosity/inquisitiveness	—	0.765	—	—	—	0.701
16	Positive professional attitude	—	0.736	—	—	—	0.634
Factor 3 — Interpersonal and communication skills (Items 7–10)							
7	Communication/listening skills	—	—	0.72	—	—	0.585
8	Empathy, non-judgemental attitude	—	—	0.499	—	—	0.397
9	Teamwork	—	—	0.613	—	—	0.405
10	Cultural/religious respect	—	—	0.61	—	—	0.423
Factor 4 — Professional competencies and self-management (Items 20–24)							
20	Time management/punctuality	—	—	—	0.617	—	0.535
21	Leadership quality	—	—	—	0.819	—	0.895
22	Body language management	—	—	—	0.886	—	0.942
23	Self-care/wellbeing	—	—	—	0.704	—	0.539
24	Digital/health literacy	—	—	—	0.876	—	0.904
Factor 5 — Reflective practice and lifelong learning (Items 17–19)							
17	Recognition of own limits	—	—	—	—	0.676	0.743
18	Coping with ambiguity	—	—	—	—	0.69	0.73
19	Lifelong learning commitment	—	—	—	—	0.765	0.76

Note. Factor loadings < 0.40 are suppressed and shown as —. h² = communality. Extraction method: maximum likelihood; rotation: varimax with Kaiser normalisation.

3.4 Test-Retest Reliability (Intraclass Correlation Coefficient)

Domain-level ICCs (two-way mixed model, absolute agreement, two-week interval) were: Domain 1 = 0.781; Domain 2 = 0.762; Domain 3 = 0.795; Domain 4 = 0.782; Domain 5 = 0.765. The overall scale ICC was 0.834. All domain and overall ICCs exceeded the pre-specified threshold of 0.70.

3.5 Vignette Validation: Expert Inter-Rater Agreement (Fleiss' Kappa)

Six of the ten vignette questions achieved unanimous agreement ($P_i = 1.00$). The remaining four questions (V2.Q1, V4.Q1, V5.Q1, V5.Q2) recorded $P_i = 0.80$, each with one dissenting expert. Mean observed agreement was 0.900; mean expected chance agreement was 0.436; overall Fleiss' kappa was 0.823, exceeding the pre-specified threshold of $\text{kappa} > 0.61$. For V3.Q1, collapsing the two pre-specified acceptable answers (B and D) into a single category yielded $P_i = 1.00$.

Table 3. Expert inter-rater agreement on the vignette scoring key (n = 10 experts, 10 questions)

Question	Expert selections	Item agreement (P_i)	Consensus level	Correct answer
V1.Q1	C (10/10)	1.000	Perfect	C
V1.Q2	B (10/10)	1.000	Perfect	B
V2.Q1	C (9/10), B (1/10)	0.800	Strong	C
V2.Q2	B (10/10)	1.000	Perfect	B
V3.Q1	B (9/10), D (1/10)	1.000*	Perfect (collapsed)	B/D
V3.Q2	B (10/10)	1.000	Perfect	B
V4.Q1	B (9/10), E (1/10)	0.800	Strong	B
V4.Q2	B (10/10)	1.000	Perfect	B
V5.Q1	E (9/10), B (1/10)	0.800	Strong	E
V5.Q2	E (9/10), C (1/10)	0.800	Strong	E
Overall Fleiss' kappa		0.823	Almost perfect	Threshold met (>0.61)

Note. *For V3.Q1, both options B and D were pre-specified as acceptable; agreement was perfect when the two options were collapsed.

3.6 Summary of Validation Results

Table 4 presents a consolidated summary of the psychometric properties examined across all six validation phases. Domain-level statistics below follow the a priori five-domain structure used

for scale administration and scoring (e.g., Domain 4 comprises Items 17–20); this grouping differs slightly from the empirically extracted factor structure reported in Table 2, where Item 20 cross-loaded onto Factor 4 (Items 21–24) rather than its hypothesised domain (Items 17–19 in Factor 5), a discrepancy discussed further in Section 4.1.

Table 4. Consolidated validation summary — NAQ scale (24 items, 5 domains)

Domain	Items	S-CVI/I-CVI	Cronbach's alpha	ICC (test-retest)	Verdict
Domain 1: Professionalism	1–6	0.93	0.946	0.781	Validated
Domain 2: Communication and interpersonal	7–10	0.93	0.855	0.762	Validated
Domain 3: Personal qualities	11–16	0.93	0.943	0.795	Validated
Domain 4: Professional adaptability	17–20	0.93	0.916	0.782	Validated
Domain 5: Modern competencies	21–24	0.93	0.926	0.765	Validated
Overall scale (24 items)	1–24	0.82 (S-CVI)	0.939	0.834	Validated

4. Discussion

4.1 Interpretation of Findings in Relation to Objectives and Literature

This study undertook a comprehensive, multi-phase psychometric validation of the Non-Academic Qualities (NAQ) scale and its companion vignette battery, designed to assess the perceived importance of non-academic physician qualities across a uniquely broad range of stakeholder groups within medical education. Taken together, the results provide strong and convergent evidence of validity and reliability across every psychometric domain examined: content validity (S-CVI = 0.82), face validity (mean Clarity Index = 0.958; mean Relevance Index = 0.936), internal consistency (overall alpha = 0.939; domain alpha = 0.855–0.946), construct validity (a clean five-factor solution explaining 66.9% of variance, closely mirroring the a priori domain structure), test-retest reliability (overall ICC = 0.834; domain ICC = 0.762–0.795), and vignette scoring key validity (Fleiss' kappa = 0.823, "almost perfect" agreement). These findings position the NAQ scale as a methodologically robust instrument suitable for deployment in the planned multi-stakeholder main study, directly fulfilling the study's stated objective of developing and

validating a psychometrically sound instrument for assessing non-academic qualities.

Content and face validity in context. The S-CVI of 0.82 exceeds the minimum acceptability threshold of 0.80 recommended by Polit and Beck (2006) and approaches the 0.90 benchmark conventionally regarded as indicative of excellent content validity. This is broadly consistent with comparable validation work in medical education; a recent validation of a theory of planned behaviour questionnaire assessing professional behaviours in medical students reported Item-level Content Validity Index values of 0.9–1 for relevance, a similar order of magnitude to the present findings. The four items achieving the highest expert consensus (I-CVI = 0.90) span both traditional (patient-centredness) and emerging (digital literacy) domains of professionalism, suggesting expert consensus extends to newer competency domains and not merely long-established ones. The exceptionally high mean Clarity Index (0.958) further indicates the scale is comprehensible to its intended lay and professional respondents alike, addressing a known limitation in non-academic attribute measurement — namely, that constructs such as empathy and emotional intelligence are

notoriously difficult to operationalise in respondent-friendly language.

Internal consistency and the treatment of Item 9. The overall Cronbach's alpha of 0.939 substantially exceeds the conventional 0.70 threshold and falls within the range George and Mallery (2003) classify as excellent. Domain-level alpha values (0.855–0.946) were uniformly strong, supporting the internal coherence of each of the five hypothesised domains. Item 9 (ability to work effectively in a team), with a corrected item-total correlation of 0.159 (below the conventional 0.20 minimum), warrants specific discussion. Restricted variance and ceiling effects are a recognised cause of attenuated item-total correlations in scales measuring near-universally endorsed constructs (Nunnally & Bernstein, 1994), and the negligible impact of item removal on overall alpha (0.939 to 0.942) confirms this is a measurement artefact rather than conceptual misfit. This was corroborated by the EFA, in which Item 9 loaded cleanly onto its hypothesised interpersonal factor (loading = 0.613) despite its weak item-total correlation — illustrating the value of triangulating multiple psychometric indices rather than relying on any single statistic to guide item retention decisions.

Construct validity and comparison with existing frameworks. The EFA yielded a five-factor solution accounting for 66.9% of total variance, with KMO = 0.847 and a highly significant Bartlett's test ($p < .001$), confirming the data were well suited to factor analysis (Kaiser, 1974). The close correspondence between the empirically derived factor structure and the a priori domain organisation, with four of five domains achieving perfect alignment, lends strong support to the construct validity of the NAQ scale. This structure maps coherently onto established professionalism frameworks, including the CanMEDS roles and the General Medical Council's Good Medical Practice domains, while extending them through contemporary constructs — ethical decision-making, accountability, self-care and wellbeing, and digital and health literacy — that were largely absent from earlier instruments such as

the Lambe and Bristow (2010) Delphi-derived attribute list. The minor cross-loading tendency observed for Item 20 (time management and punctuality) toward the professional competencies factor rather than its hypothesised professional adaptability domain is conceptually plausible. Time management arguably reflects both adaptive capacity and active self-management, and its retention within the original domain on theoretical grounds is consistent with recommendations to balance empirical and theoretical considerations in scale refinement (Nunnally & Bernstein, 1994).

Test-retest reliability. The overall test-retest ICC of 0.834, with domain-level values ranging from 0.762 to 0.795, indicates good temporal stability per the Koo and Li (2016) framework, with all values exceeding the pre-specified 0.70 threshold. Attitudinal and opinion-based items, such as those comprising the NAQ scale, are generally expected to show somewhat greater temporal variability than purely factual or behavioural items, given their susceptibility to transient contextual and affective influences (Streiner et al., 2015); the present ICC values, while good rather than excellent, are therefore well within the expected and acceptable range for an instrument of this nature.

Vignette validation as a methodological contribution. The near-unanimous expert agreement on the vignette scoring key (Fleiss' kappa = 0.823, six of ten questions achieving perfect 10/10 agreement) substantially exceeded the pre-specified threshold of kappa > 0.61. It also approached the inter-rater reliability levels reported for the Medical Judgment Vignettes interview used in medical school admissions research, where scored interviews achieved kappa = 0.96 and three-station generalizability coefficients of up to 0.81. This level of agreement is notable given that vignette-based judgement tasks are inherently more susceptible to rater disagreement than straightforward relevance ratings, since they require experts to weigh competing professional considerations rather than simply endorsing a stated item. The perfect agreement obtained for V3.Q1 once the two pre-specified acceptable response options

were collapsed illustrates that apparent rater disagreement in vignette-based instruments may sometimes reflect intentional, theoretically justified response flexibility rather than genuine measurement error. This distinction has direct relevance for how scoring keys are constructed and reported in future situational judgement instruments. To our knowledge, this is among the first studies to formally validate a vignette battery using Fleiss' kappa as a companion behavioural-validation layer alongside a structured Likert importance scale within the same instrument, rather than deploying vignette-based assessment as a stand-alone selection tool.

4.2 Implications for Theory, Practice, and Policy

This validation study extends the existing literature in three principal respects, with implications spanning theory, assessment practice, and curricular/admissions policy.

- **Theoretical implications:** By integrating a structured, multi-domain Likert importance scale with a parallel vignette-based behavioural validation component, this study enables future triangulation between stated and applied professionalism judgements — a combination not previously reported in the non-academic physician attribute literature. This dual-layer design offers a methodological template for future instrument development in health professions education, where self-report and applied-judgement measures are typically validated separately rather than as a companion pair.

- **Practical implications for assessment:** The instrument was designed for deployment across an unusually broad stakeholder ecosystem, spanning medical students at multiple training stages, postgraduate residents, faculty, staff nurses, allied health students, and the lay public. This considerably exceeds the scope of earlier single- or dual-stakeholder studies such as Lambe and Bristow's (2010) clinician-only Delphi survey or comparative faculty-student instruments restricted to the teaching context. This breadth supports the NAQ scale's use in multi-

stakeholder curriculum evaluation, admissions research, and 360-degree professionalism assessment.

- **Policy implications for curriculum design:** The inclusion of contemporary, competency-based items — ethical decision-making, accountability, self-care, and digital and health literacy — reflects current priorities in competency-based medical education that were largely unaddressed in earlier-generation instruments. This strengthens the contemporary relevance and applicability of the NAQ scale for institutions revising curricula or admissions criteria to better align with modern professionalism frameworks (e.g., CanMEDS, Good Medical Practice).

4.3 Limitations

Several limitations warrant consideration:

- **Sample size and generalisability:** The expert panel ($n = 10$) and pilot sample ($n = 50$, with the same group also completing the test-retest administration in Phase V), while consistent with established minimum recommendations for content validation and pilot reliability studies (Lynn, 1986), are relatively modest in size. Replication with larger and more geographically diverse expert and pilot samples would strengthen confidence in the generalisability of these psychometric estimates.

- **Test-retest interval:** The two-week test-retest interval, while standard practice, cannot fully exclude the possibility of residual memory effects on individual item responses, although the absence of systematic score drift between administrations is reassuring in this regard.

- **Single-institution scope:** As a single-institution validation study, the demographic and professional composition of the expert and pilot samples may not fully represent the diversity of stakeholder perspectives intended for the main study. The forthcoming multi-site, multi-stakeholder administration will provide an opportunity to further examine measurement invariance across these diverse respondent groups.

5. Conclusion

5.1 Summary of Key Findings

This study successfully developed and psychometrically validated the 24-item Non-Academic Qualities (NAQ) scale and its companion five-vignette battery through a rigorous six-phase design. The instrument demonstrated strong, convergent evidence of validity and reliability across every domain examined:

- **Content validity:** All 24 items achieved $I\text{-}CVI \geq 0.80$, with an overall $S\text{-}CVI$ of 0.82.
- **Face validity:** Pilot testing confirmed excellent clarity and relevance (mean Clarity Index = 0.958; mean Relevance Index = 0.936).
- **Internal consistency:** Overall Cronbach's alpha was 0.939, with domain-level alpha ranging from 0.855 to 0.946.
- **Construct validity:** Exploratory factor analysis yielded a coherent five-factor solution explaining 66.9% of variance, closely mirroring the a priori domain structure.
- **Test-retest reliability:** Overall ICC was 0.834 (domain range 0.762–0.795), indicating good temporal stability.
- **Vignette scoring key validity:** Expert inter-rater agreement was almost perfect (Fleiss' kappa = 0.823).

Taken together, the Non-Academic Qualities scale and its companion vignette battery demonstrate strong, convergent evidence of content validity, face validity, internal consistency, construct validity, test-retest reliability, and expert-validated vignette scoring, supporting their use as a psychometrically sound instrument for assessing non-academic qualities of MBBS graduates across a uniquely broad range of stakeholder perspectives. These findings provide the empirical foundation for the planned main-phase study, in which the validated NAQ scale and vignette battery will be administered across twelve stakeholder groups to examine importance–performance gaps and inform evidence-based curriculum development in medical education.

5.2 Recommendations and Future Research Directions

Building on these validation findings, several directions are recommended for future work:

1. **Main-phase multi-stakeholder deployment.** The immediate next step is administration of the validated NAQ scale and vignette battery across the planned twelve stakeholder groups (medical students at multiple training stages, postgraduate residents, faculty, staff nurses, allied health students, and the lay public) to examine importance–performance gaps and generate stakeholder-specific insights for curriculum reform.
2. **Measurement invariance testing.** Given the single-institution scope of this validation study, future research should employ confirmatory factor analysis and measurement invariance testing (e.g., across stakeholder groups, training stages, or institutions) to confirm that the five-factor structure holds equivalently across the diverse populations for which the instrument is intended.
3. **Larger and more diverse samples.** Replication with larger, multi-site expert panels and pilot samples would strengthen confidence in the generalisability of the psychometric estimates reported here, particularly the content validity and clarity indices derived from relatively modest samples ($n = 10$ and $n = 50$, respectively).
4. **Predictive and criterion validity.** Future studies should examine the NAQ scale's relationship to external, real-world outcomes — such as supervisor-rated professionalism, disciplinary or fitness-to-practise records, or patient satisfaction measures — to establish criterion-related and predictive validity beyond the content, construct, and reliability evidence established here.
5. **Longitudinal and curricular impact studies.** Given the instrument's grounding in contemporary competency-based frameworks (e.g., digital literacy, self-care, ethical decision-making), longitudinal designs tracking NAQ scores across training stages could clarify how

non-academic qualities develop over time and whether targeted curricular interventions produce measurable gains.

6. Refinement of the vignette battery.

While the vignette scoring key achieved almost

7. scales to broader use.

perfect expert agreement, expanding the vignette battery to cover additional professionalism domains, or piloting alternative response formats, may further strengthen the applied-judgement component of the instrument as it

References

- Albanese, M. A., Snow, M. H., Skochelak, S. E., Huggett, K. N., & Farrell, P. M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine*, 78(3), 313–321.
- Artino, A. R., La Rochelle, J. S., Dezee, K. J., & Gehlbach, H. (2014). Developing questionnaires for educational research: AMEE Guide No. 87. *Medical Teacher*, 36(6), 463–474.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Frank, J. R., Snell, L., & Sherbino, J. (Eds.). (2015). *CanMEDS 2015 physician competency framework*. Royal College of Physicians and Surgeons of Canada.
- General Medical Council. (2013). *Good medical practice*. General Medical Council.
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*, 11.0 update (4th ed.). Allyn and Bacon.
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Lambe, P., & Bristow, D. (2010). What are the most important non-academic attributes of good doctors? A Delphi survey of clinicians. *Medical Teacher*, 32(8), e347–e354.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–385.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education*, 50(1), 36–60.
- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.
- Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Acknowledgements

The authors wish to thank the ten expert panel members and the fifty pilot-phase respondents for their time and thoughtful participation in this validation study. Their contributions to the item review, clarity and relevance ratings, and vignette scoring key are gratefully acknowledged. This research received no specific grant or funding from any funding agency in the public, commercial, or not-for-profit sectors.

Author Declaration

The authors confirm that this work is original and has not been submitted elsewhere for publication, in whole or in part, nor is it currently under review at another journal. The authors declare no conflicts of interest in relation to this study. This study was approved by the Institutional Human Ethics Committee, Chettinad Academy of Research and Education (CARE IHEC-II) (Approval No. IHEC-II/0209/22, approved 30.06.2022). The ethical

issues of the research are explained in the paper and in the accompanying Materials & Methods file submitted separately.

Supplementary Material (for online publication only, linked from the main text)

Supplementary Table S1. Full wording of the 24 Non-Academic Qualities (NAQ) scale items

Item	Full item wording
1	Recognition that patient care is the primary concern of a doctor
2	Honest, trustworthy and acting with integrity
3	Ethical decision-making
4	Accountability and responsibility — owning clinical decisions and errors
5	Respecting patient privacy and confidentiality
6	Professional dress code and appearance
7	Good communication and listening skills
8	Empathy and non-judgemental attitude toward patients
9	Ability to work effectively in a team
10	Respect for culture, religion and community practices
11	Motivation and commitment to the medical profession
12	Resilience — ability to recover from failure and setback
13	Flexibility and adaptability
14	Creativity and innovative thinking
15	Curiosity and inquisitiveness
16	Positive attitude and professional behaviour
17	Recognition of one's own limits and those of others
18	Ability to cope with ambiguity, change, complexity and uncertainty
19	Commitment to lifelong learning, competence and performance development
20	Time management and punctuality
21	Leadership quality and ability to guide others
22	Body language management and non-verbal communication
23	Self-care and personal wellbeing
24	Digital and health literacy