Indonesian Journal of Education Research (IJoER)
Vol. 7, No. 5, October 2026, pp. 685~700
ISSN: 2722-1326, DOI: 10.37251/ijoer.v7i5.3787

What the Content Validity Index Does Not Show: Rater Severity and Cut-Point Sensitivity in a Two-Round Delphi Validation and Implications for Instrument Validation in Education

Nguyen Dinh Tho[1], Nguyen Phuong Chi[2], Pham Tho Hoan[3]
[1] Institute of Digital Technology, Thu Dau Mot University, Ho Chi Minh, Viet Nam
[2,3] School of Mathematics and Computer Science, Hanoi National University of Education, Hanoi, Viet Nam

Article Info

Article history:

Received Aug 4, 2026
Revised Sep 13, 2026
Accepted Oct 26, 2026
Online First Oct 28, 2026

Keywords:

Ceiling Effect
Competence Framework
Content Validity Index
Delphi Method
Rater Severity

ABSTRACT

Purpose of the study: This study examines what the content validity index conceals rather than what it reports. Using a two round Delphi validation, it asks how far rater severity varies within one expert panel, how sensitive the consensus conclusion is to the dichotomisation point, and which findings survive across four index specifications.

Methodology: The complete Round 2 rating matrix of a two round Delphi validation of a competence framework was re-analysed (20 experts, 169 items, 3,380 ratings). Analyses in base R comprised per rater severity summaries, cut point and threshold sensitivity, a leave one expert out jackknife, four index specifications, pairwise Spearman correlations against a permutation null model, and non-parametric facet comparisons.

Main Findings: Expert mean ratings spanned 3.94 to 4.99 and top category use spanned 1.2% to 99.4%, yet the proportion of ratings at or above 4, the only quantity the index uses, spanned just 91.1% to 100%. Omitting any single expert left all 169 items in consensus, whereas restricting relevance to the top category cut qualifying items to 16. Aiken's V and the modified kappa left the conclusion intact. Rank agreement was real but minimal.

Novelty/Originality of this study: Validation reports vary the panel and almost never vary the index specification. This study varies both on one dataset, shows that fragility is confined to stricter dichotomisation while indices using the full ordinal scale remain stable, settles by permutation whether weak rank agreement reflects disagreement or attenuation, and derives a diagnostic set requiring no new data. For education researchers and developers of assessment instruments, the practical implication is that a content validity index should be reported together with the rating distribution, per rater summaries and at least one index that uses the full scale before an instrument is used for curriculum, assessment or certification decisions.

This is an open access article under the CC BY license

Corresponding Author:
Nguyen Dinh Tho
Institute of Digital Technology, Thu Dau Mot University, 06 Tran Van On, Phu Loi, Ho Chi Minh City, Viet Nam
Email: thond@tdmu.edu.vn

1. INTRODUCTION

When an instrument is appraised by experts rather than by test takers, the content validity index has become the number that carries the argument. A panel rates each item for relevance, the proportion rating it in the upper categories is computed as an item level index (I-CVI), the average across items is reported as a scale

level index (S-CVI/Ave), and values above conventional thresholds are taken to license the instrument [1]–[3]. The procedure is attractive because it is simple, transparent and easy to compare across studies, and reviews of applied practice confirm how widely it has been adopted [4], [5]. Its simplicity, however, rests on a step that is rarely interrogated: an ordinal judgement on a four or five point scale is collapsed into a binary one, and everything the index subsequently reports is a function of where that collapse is made.

Two consequences follow, and neither is routinely examined in applied validation reports. The first concerns raters. Because the index counts only whether a rating falls above the cut point, two experts who use the scale in radically different ways, one reserving the top category for exceptional items and another using it habitually, contribute identically to the index provided both stay above the cut. Severity is among the best documented and most consequential of rater effects [6]–[10], yet the differences that a measurement model would treat as a parameter to be estimated are, for the content validity index, invisible by construction. The second concerns the cut point itself. Applied studies commonly report sensitivity analyses that vary the panel, dropping inattentive or outlying raters, but almost never report what happens if the dichotomisation point moves, even though the latter is a researcher decision of exactly the same kind and, as we show, of far greater consequence.

Alternatives to dichotomisation have existed for decades and are seldom reported alongside the index. The content validity ratio [11] and Aiken's V [12], [13] were both designed for ordinal expert ratings, and V in particular uses the entire response range rather than collapsing it. Polit, Beck and Owen [2] proposed a modified kappa that corrects the index for chance agreement, a correction that applied studies citing their paper frequently omit. Reviews of content validation methods [4], [14]–[16] and of validity theory more broadly [17]–[20] converge on the same point: content validity is a body of evidence rather than a single coefficient. In the Delphi literature the parallel debate concerns how consensus is defined at all, with reviews showing that thresholds are chosen inconsistently and often left unjustified [21]–[28].

This matters beyond methodological tidiness. Validation reports are used to justify instruments that then determine what is taught and assessed. If the reported evidence is largely a function of an arbitrary collapse, then two things are at stake: the credibility of the warrant, and, more practically, the temptation to read item level rankings out of the index in order to refine the instrument. The second use is the more common and the less defensible.

The concern is practical in education, where instruments of this kind are produced continually. Competence frameworks, teacher competence instruments, scoring rubrics and statements of intended learning outcomes are routinely put before expert panels before they are adopted, and the resulting index is then cited as evidence of content validity in curriculum documents, program self-evaluation reports and accreditation files. Current validity theory treats such evidence as one strand in an argument for a particular interpretation and use of scores rather than as a property of the instrument [29], [30], and the joint standards for educational and psychological testing call for the procedures used to obtain expert judgements to be described in full, including how the experts were selected and how their ratings were elicited [31], [32]. Computing education adds a further pressure. International curriculum recommendations have moved from lists of knowledge areas towards statements of competence [33]–[35], and those statements have to be turned into assessable criteria and rubrics before they can be used to grade student work [36], [37]. A framework validated by a Delphi panel, a method with a long history in educational decision making [38], can therefore pass almost directly into the outcome statements of a degree program and the rubrics used to mark student projects. If the validity evidence rests on a convention that no one inspects, an institution can document compliance with a quality requirement while the instrument itself says little about which of its elements matter most.

Earlier work has examined parts of this problem separately. The index has been criticised for ignoring chance agreement [2], [39]; reviews of Delphi studies have shown that consensus thresholds are chosen inconsistently [21], [22]; and the statistical literature has documented what is lost when ordinal or continuous scores are dichotomised [40], [41]. These critiques have mostly been argued from first principles or from simulated data, and each addresses one analytic decision at a time. We have not found a study that takes one complete expert rating matrix and varies the rater set, the dichotomisation point, the decision threshold and the index specification on the same items and the same raters, so that the influence of each decision can be compared directly. Without that comparison an applied researcher cannot tell which of the possible sensitivity checks deserves to be reported. The question is also pressing in practice. Competence based curricula in computing [33], [34] are producing a growing number of frameworks validated by expert panels, and each one reported with an I-CVI table alone adds to a body of evidence whose dependence on a single analytic convention has not been measured. The primary objective of this study is therefore to determine, on one fully disclosed Delphi dataset, how much of a content validity conclusion is carried by the expert judgements and how much by the dichotomisation applied to them, and to derive from that comparison a minimum diagnostic set that education researchers can report without collecting new data.

We examine both issues empirically using the complete rating matrix of a two round Delphi validation that we conducted ourselves. The panel judged the framework defensible as a whole: all 169 rated items met

three consensus criteria specified in advance, a verdict that follows correctly from those criteria and from established guidance [1], [3]. We do not revisit it. We ask a different question: given the same data, what else can be recovered, and what does the index hide? Three research questions structure the analysis.

RQ1. How widely does rater severity vary within a single expert panel, and how much of that variation is visible to the content validity index?

RQ2. How sensitive is the consensus conclusion to the dichotomisation point and to the decision threshold, compared with its sensitivity to panel composition?

RQ3. Which conclusions are invariant across four index specifications, and does the panel agree on the ordering of items once the ceiling effect is accounted for?

A note on terminology. We refer throughout to problem-solving competence in programming rather than to programming problem-solving competence, because the prepositional form locates the domain of activity rather than narrowing the construct to problems of the programming exercise type. Both forms are current in the literature and nothing in the present analysis depends on the distinction.

2. RESEARCH METHOD

2.1. Research Type and Design

This is a quantitative, secondary methodological study. It re-analyses a complete set of expert ratings collected in an earlier Delphi survey and gathers no new data. The Delphi technique asks a fixed panel of experts to rate the same material in successive rounds, with feedback between rounds and anonymity among panellists, so that a group judgement can be formed without face to face discussion [27], [28], [42], [43]. The present study does not test a hypothesis about the framework the panel judged. It treats the rating matrix as fixed and asks how the conclusion drawn from it changes when the analytic conventions applied to it change. In that sense the design is a sensitivity analysis: the data are held constant while the rater set, the dichotomisation point, the decision threshold and the index specification are varied. Where they apply to a re-analysis, the reporting items recommended for Delphi studies [26], [44] are addressed in Sections 2.2 to 2.6.

2.2. Research Participants

The data are the complete Round 2 rating matrix of a two round Delphi study conducted between May and July 2026 with a panel of 23 experts fixed before the first round and drawn purposively from three groups: programming instruction and information technology education, software industry recruitment and development, and educational measurement. Round 1 returned 21 responses and Round 2 returned 20 (87.0% of the fixed panel), of which 19 came from experts who had also responded in Round 1. Inclusion required a postgraduate qualification and at least five years of relevant professional experience (Appendix A.3). The one expert who responded only in Round 2 received the Round 1 summary materials before rating.

2.3. Instrument and Data Collection

Each respondent rated 169 items on five point scales, distributed across four blocks. Block A comprised five competence components judged on four facets (20 items). Block B comprised 21 criteria judged on the same four facets, namely relevance, clarity, necessity and measurability (84 items). Block C comprised the four developmental levels judged on three facets, clarity of description, observability and curriculum fit (12 items), together with three judgements of distinctiveness between adjacent levels, giving 15 items in total; distinctiveness yields three judgements rather than four because four ordered levels form only three adjacent boundaries. Block D comprised 25 sample behavioural indicators judged on two facets (50 items). An attention check embedded within Block C was excluded from all analyses. Sixteen of the 20 respondents answered it as instructed and four left it blank; all 20 respondents were retained in the analysis. The matrix therefore contains $20 + 84 + 15 + 50 = 169$ items and 3,380 ratings with no missing values.

The instrument was a structured rating questionnaire built directly from the framework in Appendix A, with one rating for each combination of element and facet. Round 1 results were summarised and returned to the panel before Round 2 was administered, which is the controlled feedback that distinguishes a Delphi survey from a repeated questionnaire [42], [44]. The consensus criteria against which the ratings were judged were fixed before Round 1, are reproduced in Appendix A.4, and define the primary analysis in Section 2.5; they were not altered after the data were seen. Only the Round 2 matrix is analysed here, because it is the matrix on which the consensus verdict rests.

2.4. Related Manuscripts, Ethics and Data Availability

This is a re-analysis of data the authors collected for a different purpose. The Delphi study is a data collection exercise, not a prior publication, and no new data were gathered for the present paper. Appendix A sets out the framework, the appraisal facets, the composition of the 169 judged items, the panel and the consensus criteria fixed before data collection, so that every claim made below can be checked against this

article and its open dataset alone. The first row of Table 2 is not a figure taken from another source: it is computed from the same matrix by the script accompanying this paper, and serves as the reference point against which the sensitivity analyses are read. In the interest of transparency we record that two further manuscripts by the same authors draw on the same dataset. The first reports the development of the framework and the overall content validity verdict and is at present under review at another journal; no claim in this paper depends on it, and because a manuscript under review cannot be retrieved by a reader we disclose it here rather than cite it. The second has been accepted for publication and develops a many facet Rasch design for rubric based assessment; it is cited at Section 3.9 [45] as the place where that design is worked out, while the recommendation made there rests on the published measurement literature cited in that section rather than on it. The boundary of contribution is that the present paper asks how far a content validity conclusion depends on the analytic conventions applied to it, which neither of the other manuscripts addresses.

Participation was voluntary and by invitation. Every expert gave informed consent before the Round 1 questionnaire was released, having been told the purpose of the study, the intended use of the ratings and the anonymisation procedure; consent covered both rounds and secondary methodological analyses of the kind reported here. Identifying information was collected solely to link responses across rounds, was replaced by arbitrary identifiers immediately after linkage, and appears in no analysis file, so that no individual expert is identifiable in this report. The de-identified Round 2 rating matrix and the complete base R script that reproduces every table, figure and statistic reported below are openly available at [repository DOI or URL].

2.5. Data analysis techniques

We computed I-CVI as the proportion of experts rating an item in the upper categories and S-CVI/Ave as the mean I-CVI across items, following the rule fixed before data collection in treating ratings of 4 or 5 as indicating relevance and applying a decision threshold of 0.78 for panels larger than eight members [1], [3]. Eight analyses followed. Table 1 maps each analysis to the research question it serves and states, in non-technical terms, what it is designed to show.

a) Response distribution.
The marginal distribution of all 3,380 ratings, to characterise the extent of ceiling.

b) Rater severity.
For each expert, the mean rating across the 169 items, the within expert standard deviation, the proportion of top category ratings, and the proportion of ratings at or above 4. The last quantity is the only one the index uses; comparing its range with the range of the others quantifies how much severity variation the index absorbs. Near constant use of one or two categories is treated in the survey literature as a response style rather than as a judgement about content [46], [47], which is why it is described separately from severity.

c) Cut point sensitivity.
S-CVI/Ave, the number of items meeting the decision threshold, and the minimum I-CVI, recomputed under a stricter dichotomisation (top category only) and with and without the three experts whose within expert standard deviation fell below 0.20. Separately, the number of items meeting decision thresholds of 0.78, 0.80, 0.90 and 1.00 while holding the dichotomisation point fixed. The 0.20 exclusion rule was applied post hoc as a descriptive check and is reported as such. Dichotomising an ordinal scale discards information, and how much is lost depends on where the cut is placed [40], [41]; this is the reason the cut point is varied here rather than assumed.

d) Panel composition sensitivity (jackknife).
The same three quantities recomputed 20 times, each omitting one expert, to compare the influence of individual raters with the influence of the cut point. The leave one out design follows the jackknife logic described by Efron and Tibshirani [48].

e) Alternative index specifications.
Two established alternatives were computed on the same matrix. The modified kappa of Polit, Beck and Owen [2], $k^* = (\text{I-CVI} - pc) / (1 - pc)$, where pc is the binomial probability of chance agreement, corrects the index for chance while retaining the dichotomisation. Aiken's V [12], [13], $V = (M - lo) / (c - 1)$ with M the item mean, $lo = 1$ and $c = 5$, uses the full ordinal range and does not dichotomise at all. The chance correction builds on the approach proposed by Wynd, Schmidt and Schaefer [39]. A score confidence interval for V is available [49] but was not needed for the descriptive use made of V here. Reporting the conclusion under four specifications rather than two is what allows RQ3 to be answered.

f) Rank agreement and a permutation null model.
Spearman correlations were computed between every pair of experts across the 169 items (190 pairs). Because 98% of ratings fall in two adjacent categories, these coefficients are attenuated by ties, so the observed mean was referred to a null distribution generated by independently permuting each expert's own rating vector 2,000 times [50]. Permutation preserves every expert's marginal distribution, and therefore the ceiling, while

destroying any shared ordering; the comparison separates genuine agreement from range restriction, which inspection of the coefficients alone cannot do.

g) Stability of the substantive signal.

Facet level results were recomputed under the alternative specifications, and facet level mean ratings compared using Wilcoxon signed rank tests [51] with the 21 criteria as the unit of analysis and paired standardised differences as effect sizes [52], together with a Friedman test [53] using the 20 experts as the unit. The three pairwise comparisons involving measurability were Bonferroni corrected [54].

h) Severity and expert profile.

Expert severity was compared across highest qualification, years of experience and professional group using Kruskal-Wallis tests [55], linking Round 2 respondents to the demographic information they had supplied in Round 1 (19 of 20 could be linked).

All analyses were conducted in base R [56] without additional packages. Quantiles use the linear interpolation definition, no analysis involved imputation, and the permutation seed is fixed in the script so that every figure and statistic reported below is exactly reproducible.

Table 1. Analyses Mapped to the Research Questions They Serve

Analysis	Research question	What the analysis asks, in plain terms	Reported in
(a) Response distribution	Context for RQ1 to RQ3	How much of the five point scale did the panel actually use? If nearly all ratings sit in the top two categories, every index that follows is computed on a compressed scale.	Section 3.1
(b) Rater severity	RQ1	Do some experts rate systematically higher than others, and does the index register the difference?	Section 3.2; Figures 2 and 3
(c) Cut point and threshold sensitivity	RQ2	Would the verdict change if only the top rating counted as relevant, or if the pass mark were raised?	Section 3.3; Table 2
(d) Jackknife	RQ2	Does the verdict depend on any single expert?	Section 3.3; Figure 4
(e) Alternative indices	RQ3	Does the verdict hold when chance agreement is removed, or when the whole scale is used instead of a yes or no split?	Section 3.4; Table 3
(f) Rank agreement and permutation null	RQ3	Do experts agree on which items are better than others once the crowding of ratings at the top is allowed for?	Section 3.5; Figures 5 and 6
(g) Facet level stability	RQ3	Is there a pattern, such as one facet always rated lowest, that survives every index?	Section 3.6; Table 4; Figure 7
(h) Severity and expert profile	RQ1	Can lenient or severe raters be identified in advance from qualification or experience?	Section 3.7; Figure 8

2.6. Research Procedure

The study proceeded in five stages, summarised in Figure 1. Before data collection, the framework, the four appraisal facets and the consensus criteria were fixed (Appendix A). Between May and July 2026 the two Delphi rounds were administered to the fixed panel of 23 experts, returning 21 and 20 responses. The Round 2 matrix was then de-identified, the attention check was removed, and the 20 by 169 rating matrix was assembled. The primary consensus analysis was next reproduced from that matrix to serve as the reference point (first row of Table 2). Finally, the eight analyses described in Section 2.5 were run in base R with a fixed permutation seed, and each result was read against the reference point to answer RQ1 to RQ3.

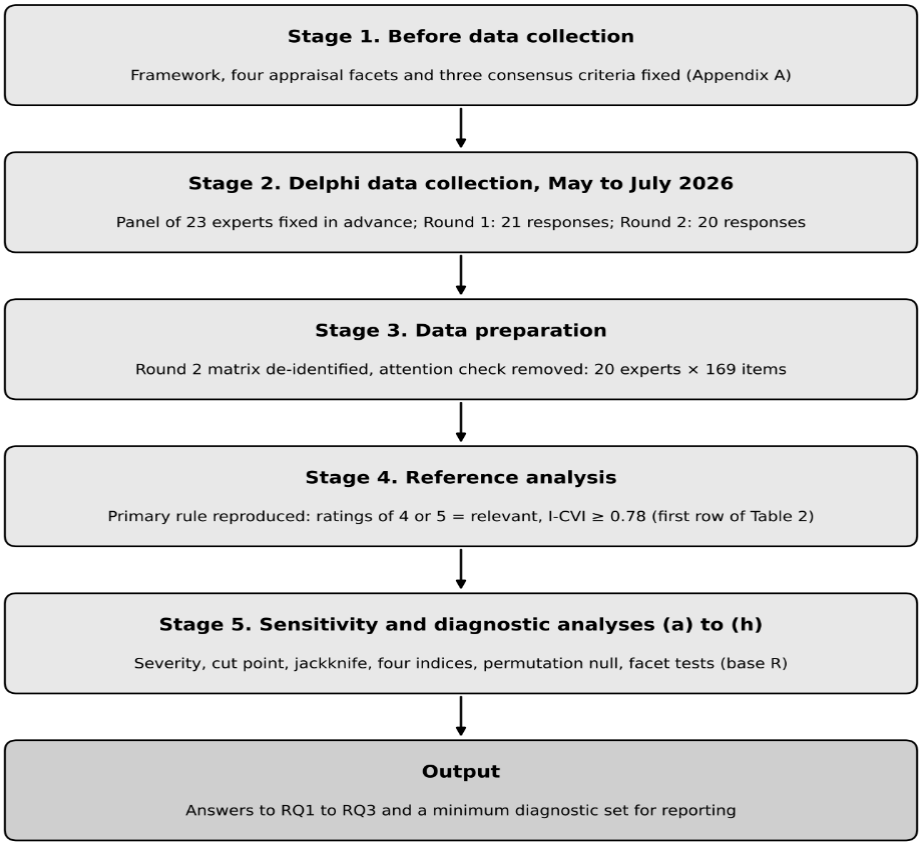


Figure 1. Research Procedure, From the Design Decisions Fixed Before Data Collection to the Diagnostic Analyses

3. RESULTS AND DISCUSSION

3.1. The Response Distribution Is Heavily Bounded

Of the 3,380 ratings, 60.5% fell in the top category and 37.5% in the next, so 98.0% of all judgements were at or above the cut point used to define relevance. Only 2.0% were at the mid point and not a single rating fell in the lowest two categories. The scale was, in practice, used as a two category instrument, and the two categories in use both count as relevance under the conventional rule. This is the structural condition under which the remaining findings should be read, and it is precisely the condition that a marginal distribution makes visible in one line and that an I-CVI table conceals entirely.

3.2. Rater Severity Varies Widely, and the Index Absorbs Almost All of It

Expert mean ratings across the 169 items ranged from 3.94 to 4.99, a span of 1.05 scale points, with a between expert standard deviation of 0.329. The proportion of top category ratings ranged from 1.2% to 99.4%: at one extreme an expert who almost never used the top category, at the other an expert who used it almost exclusively. Three of the 20 experts had a within expert standard deviation below 0.20, meaning that their ratings were close to constant across all 169 items. We describe these as near constant response styles rather than as inattentive responding, since a uniformly severe but consistent rater produces the same statistic and the distinction cannot be drawn from the ratings alone. Figures 2 and 3 display the severity ordering and the underlying response styles.

The decisive comparison is with the quantity the index uses. While mean ratings span 1.05 points and top category use spans 98 percentage points, the proportion of ratings at or above 4 spans only 91.1% to 100%, a range of nine percentage points. Two experts whose scale use has almost nothing in common therefore make near identical contributions to every I-CVI in the study. The severity variation is not small and then discarded; it is large and rendered invisible by the dichotomisation. This answers RQ1: severity varies by more than a full scale point within a single panel, and the index recovers less than a tenth of that range. The magnitude is consistent with the severity differences documented in rater mediated assessment more generally [6], [7], [9], [10], [57], [58], with the difference that those literatures estimate severity as a parameter whereas the index absorbs it.

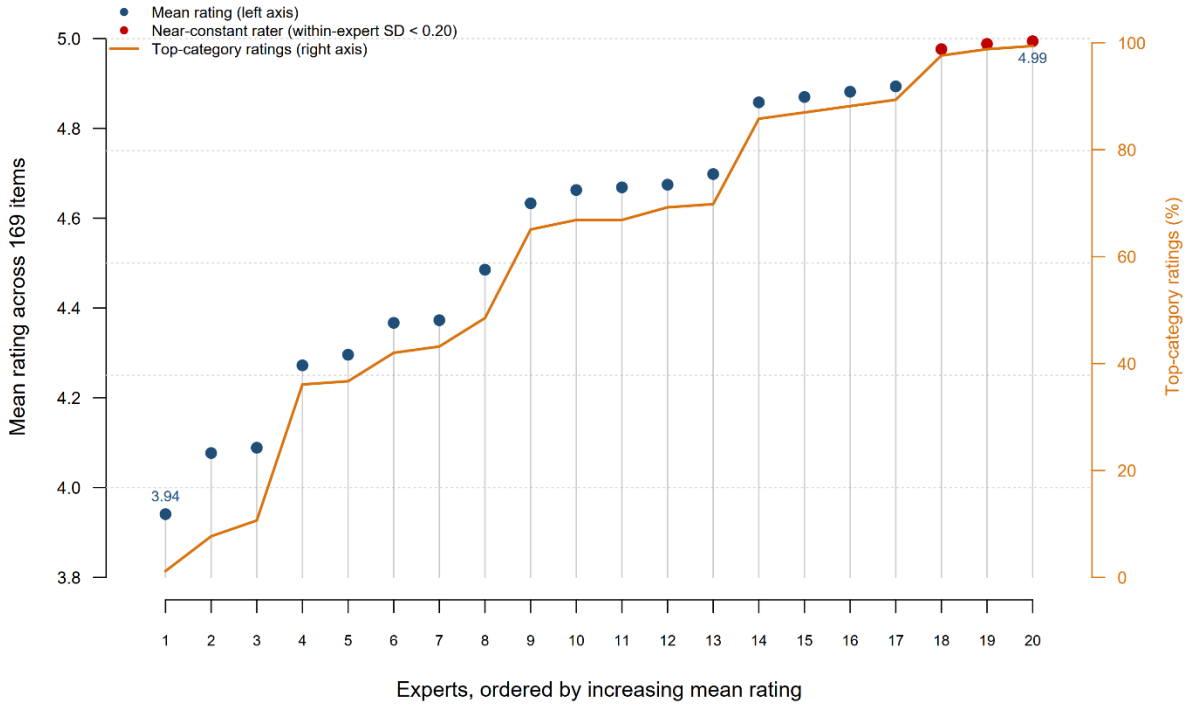


Figure 2. Rater severity across the panel: mean rating (left axis) and proportion of top category ratings (right axis), experts ordered by severity. Points in red mark the three experts whose within expert standard deviation fell below 0.20.

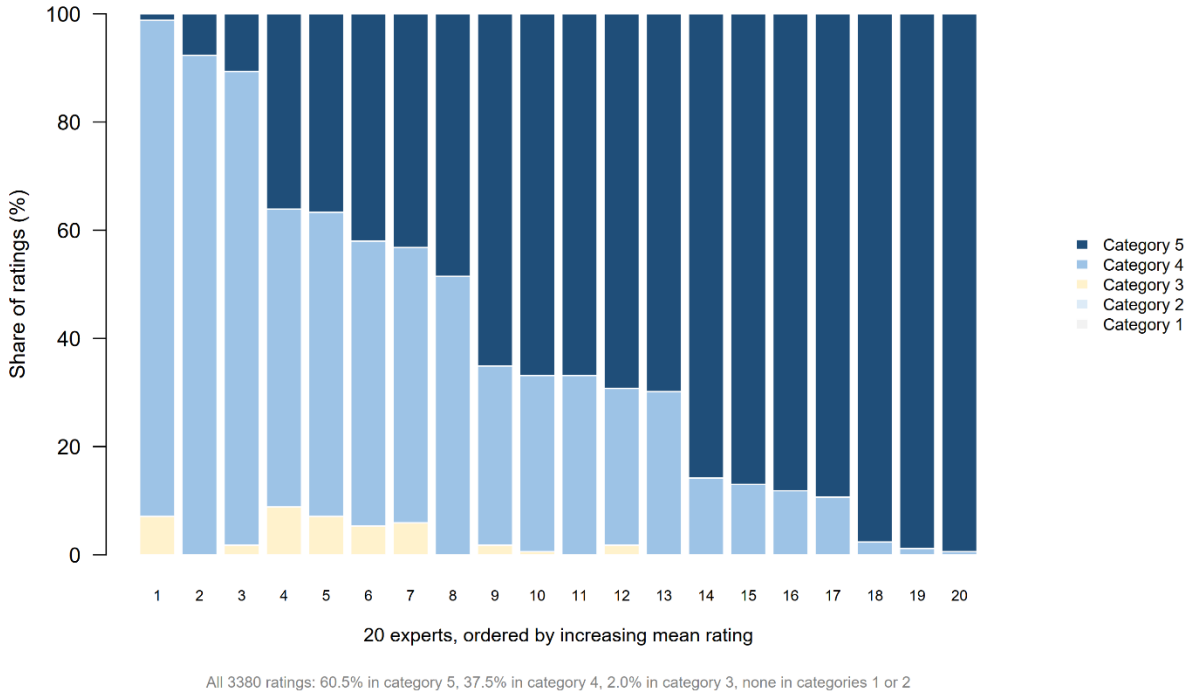


Figure 3. Response style profiles: the distribution of the five rating categories for each expert, ordered by severity.

3.3. A Robustness Asymmetry

Removing any single expert leaves the conclusion untouched. Across all 20 jackknife replications, the number of items meeting the decision threshold remained 169 without exception, the minimum I-CVI remained 0.842 in every replication, and S-CVI/Ave moved within a band of 0.0047 (from 0.9788 to 0.9835). The minimum is invariant because several items sit at 17 of 20 with different dissenting experts, so at least one item falls to 16 of 19 whichever expert is removed. By the criterion most often used to demonstrate that a Delphi result does not depend on particular panellists, this result is as robust as it could be.

Moving the dichotomisation point produces an entirely different picture (Table 2). Counting only top category ratings as indicating relevance reduces S-CVI/Ave from 0.980 to 0.605, the minimum I-CVI from 0.850 to 0.300, and the number of items meeting the 0.78 criterion from 169 to 16. Excluding the three near constant raters and applying the same stricter rule leaves only 7 items. Holding the dichotomisation fixed and raising the decision threshold instead is comparatively benign: 169 items still qualify at 0.80, 166 at 0.90, and 120 at 1.00.

Table 2. Sensitivity of the Consensus Conclusion to Analytic Choices

Analytic choice	Experts	S-CVI/Ave	Items meeting 0.78	Minimum I-CVI
Ratings of 4 or 5 count as relevant (primary analysis)	20	0.980	169 of 169	0.850
Only top category ratings count as relevant	20	0.605	16 of 169	0.300
Primary rule, three near constant raters excluded	17	0.976	169 of 169	0.824
Stricter rule, three near constant raters excluded	17	0.538	7 of 169	0.176
Primary rule, decision threshold raised to 0.90	20	0.980	166 of 169	0.850
Primary rule, decision threshold raised to 1.00	20	0.980	120 of 169	0.850

Part of this asymmetry is arithmetic rather than empirical, and saying so strengthens rather than weakens the argument. With 20 raters, omitting one can move any I-CVI by at most 1/20 = 0.05, so a jackknife can only ever perturb the index within a narrow, analytically bounded band. Moving the cut point, by contrast, relocates an entire response category across the decision boundary, and here that category holds 37.5% of all ratings. The magnitude of the cut point effect is therefore governed by the share of ratings sitting in the category immediately above the cut, a quantity that validation reports do not currently publish. What the present data add is a demonstration that in a realistic, well conducted validation this share is large enough to reverse the conclusion entirely, and that the diagnostic which would have revealed the vulnerability is a single marginal distribution.

The practical implication answers RQ2. Applied validation work routinely tests the robustness of a content validity result against panel composition, and that test is structurally easy to pass; it almost never tests robustness against the cut point, and that test is difficult to pass. Researchers have been probing the direction in which these results cannot break.

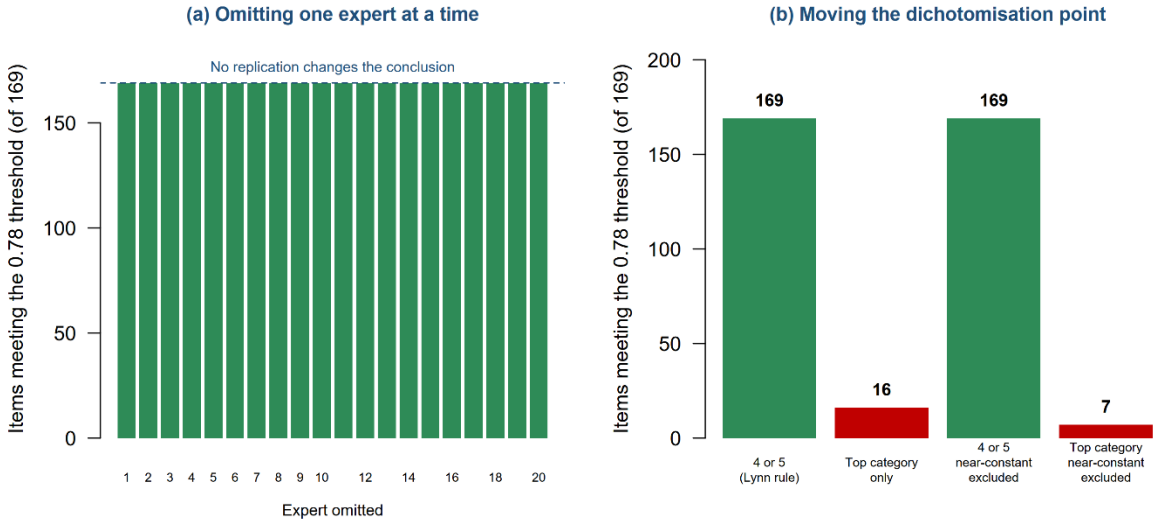


Figure 4. The robustness asymmetry: (a) number of items meeting the decision threshold under each jackknife replication; (b) the same quantity under alternative dichotomisation points.

3.4. The Fragility Belongs to Dichotomisation, Not to the Data

If the conclusion collapses when the cut point moves, a reader is entitled to ask whether the framework was ever adequately supported. Two further specifications answer that question, and they answer it in the framework's favour (Table 3). The modified kappa of Polit, Beck and Owen [2], which corrects for chance agreement while retaining the dichotomisation at 4, gave a mean of 0.980 with a minimum of 0.850, and all 169

items exceeded the 0.74 value those authors treat as excellent. Aiken's V, which uses the entire five point range and never dichotomises, gave a mean of 0.896 with a minimum of 0.788 and a maximum of 0.988. We report V descriptively rather than against a fixed pass mark, since the right tail probabilities from which a critical value is read depend on the number of raters and the number of response categories [13]; what matters here is the contrast in stability, not whether a particular threshold is cleared.

Table 3. The Conclusion Under Four Index Specifications (169 Items, 20 Experts)

Index specification	Uses full scale	Scale level value	Criterion	Items meeting criterion	Minimum
I-CVI, ratings of 4 or 5	No	0.980	≥ 0.78	169 of 169	0.850
I-CVI, top category only	No	0.605	≥ 0.78	16 of 169	0.300
Modified kappa k*, cut at 4	No	0.980	≥ 0.74	169 of 169	0.850
Aiken's V	Yes	0.896	descriptive	169 of 169	0.788

Three points follow. First, the chance correction changes almost nothing at this level of agreement: with 17 or more of 20 experts above the cut, the binomial probability of chance agreement is negligible, so k* and I-CVI coincide to three decimal places. This is worth reporting rather than assuming, because it establishes that the vulnerability identified here is not a chance agreement artefact and cannot be repaired by the correction most often recommended for it. Second, the specification that uses the full ordinal range is the one that behaves stably, and its minimum value of 0.788 sits well above the floor of the scale. Third, and consequently, the fragility documented in Section 3.3 is a property of the decision to dichotomise, and specifically of where to dichotomise, rather than a property of the underlying judgements. That distinction converts a purely negative finding into an actionable one: the remedy is not a better cut point but an index that does not require one.

3.5. Agreement on the Ordering is Real but Minimal

If the panel converged on the relative quality of items, the index would at least support ordinal use even if its absolute level were arbitrary. Across the 190 expert pairs, the mean Spearman correlation over the 169 items was +0.054 (median +0.039), the range ran from −0.192 to +0.452, and 70 pairs, 36.8%, were negatively correlated (Figure 5). Taken alone, that distribution is ambiguous. It is consistent with a panel that genuinely disagrees about relative item quality, and equally consistent with a panel that agrees but whose agreement cannot be detected because 98% of ratings occupy two adjacent categories and the rank vectors are consequently saturated with ties.

The permutation null model separates the two readings. Permuting each expert's own ratings preserves that expert's marginal distribution, and therefore the ceiling and the ties, while destroying any ordering shared with other experts. Under 2,000 such permutations the mean pairwise correlation was +0.0002 (SD 0.0057), with a 95% interval from −0.010 to +0.011. The observed value of +0.054 lies far outside that interval, approximately 9.6 null standard deviations above the null mean, with a one sided permutation $p < .001$. The panel therefore does share ordering information, and the agreement is statistically unambiguous.

Its practical magnitude is nevertheless very small. A mean rank correlation of 0.054 corresponds to well under one percent of shared rank variance between any two experts, and the large minority of negative pairs is what sampling variation around a small positive mean produces. The correct reading is thus neither of the two originally available: the panel does not disagree, and the low coefficients are not purely an artefact, but the ordering signal that survives the ceiling is far too weak to carry item level decisions. This resolves the second half of RQ3 and sharpens the practical conclusion rather than softening it. An I-CVI ranking computed from data of this shape is not merely unreliable; it is shown here to contain a real but negligible amount of information, which is a stronger reason not to act on it than uncertainty would have been.

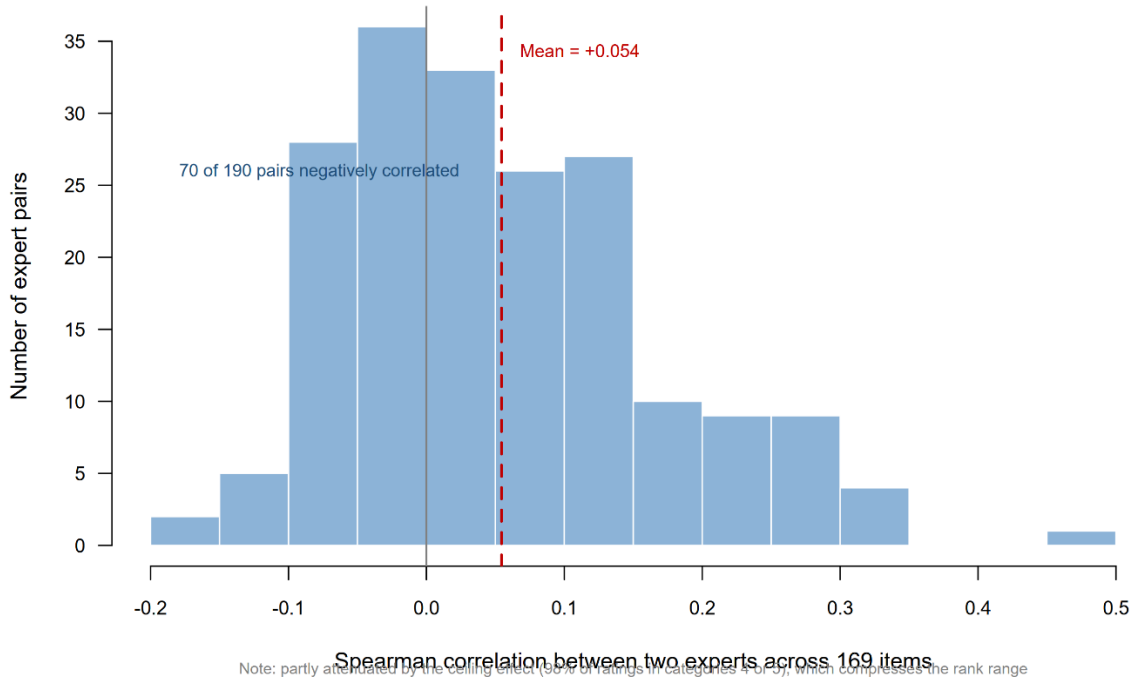


Figure 5. Distribution of pairwise Spearman correlations between experts across the 169 items.

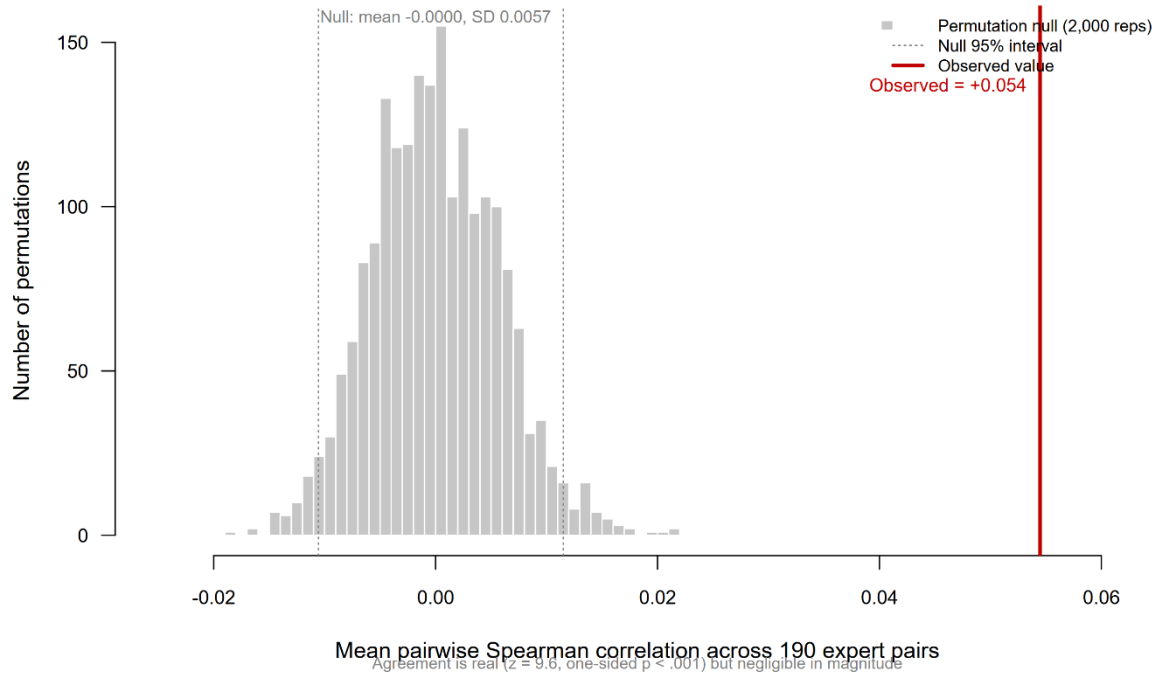


Figure 6. The observed mean pairwise correlation against the permutation null distribution (2,000 permutations preserving each expert's marginal distribution).

3.6. One Substantive Signal is Invariant Across Specifications

Not everything in the data is fragile. Among the four facets on which the 21 criteria were judged, measurability was rated lowest under every specification applied here: mean I-CVI 0.969 against 0.988 to 0.998 for the other three facets under the primary rule, 0.602 against 0.669 to 0.679 under the stricter rule, and Aiken's V 0.893 against 0.915 to 0.919 (Table 4, Figure 7). Eight of the 21 criteria had a measurability I-CVI below 1.00, against one, two and four for the other facets.

The ordering is not the product of a few extreme criteria. With the criterion as the unit of analysis (n = 21), measurability had a lower mean rating than each of the other three facets, and all three paired comparisons

remained significant at the .05 level after Bonferroni correction (adjusted p = .010, .006 and .028 against relevance, clarity and necessity), with paired effect sizes in the medium to large range (dz = 0.81, 0.95 and 0.70). A Friedman test with the 20 experts as the unit also detected a difference among the four facets, $\chi^2(3) = 8.44$, p = .038. The magnitude of those differences is not the point being made here, and we do not pursue it: what Table 4 is included to demonstrate is that the same ordering appears under the primary rule, under the stricter rule and under Aiken's V alike.

Table 4. Facet Level Results Across Specifications (21 Criteria)

Facet	I-CVI (4 or 5)	I-CVI (top only)	Aiken's V	Mean rating	Criteria with I-CVI < 1.00
Relevance	0.998	0.679	0.919	4.676	1 of 21
Necessity	0.995	0.669	0.916	4.664	2 of 21
Clarity	0.988	0.674	0.915	4.662	4 of 21
Measurability	0.969	0.602	0.893	4.571	8 of 21

The contrast with the preceding sections is instructive. Aggregated at facet level, over 21 criteria and 20 experts, the data carry a signal that survives every index specification applied here; disaggregated to individual items, they carry a signal so weak that Section 3.5 could only detect it against a null model. This is what one would expect if item level variance is dominated by rater idiosyncrasy while facet level differences reflect something the panel genuinely shares. The present paper does not attempt a substantive account of why measurability should be the weakest facet, which is a question about the construct rather than about the index. What it contributes is the demonstration that this finding, unlike the item level ordering, is invariant to the analytic choice that governs everything else in Table 2. This completes the answer to RQ3.

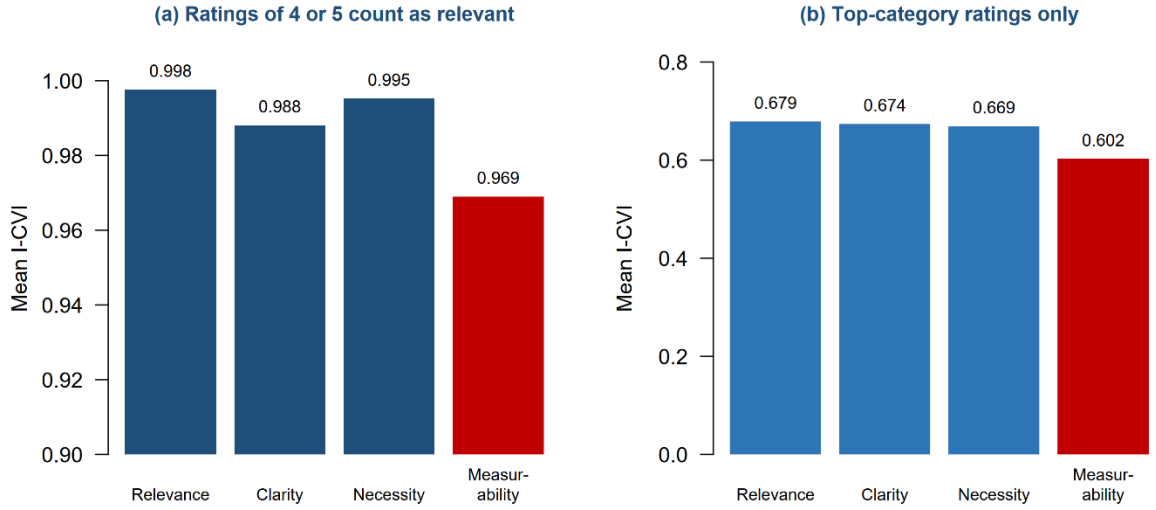


Figure 7. Measurability is the lowest rated facet under all three specifications.

3.7. Severity is not Predictable From the Expert Profile

If severity were associated with qualification or experience, it could in principle be managed at the selection stage. We found no evidence that it is. Among the 19 experts whose Round 1 demographic information could be linked to their Round 2 ratings, Kruskal-Wallis tests detected no difference in severity by highest qualification (H(2) = 1.54, p = .464) or by years of experience (H(2) = 1.99, p = .369); the professional group comparison could not be tested because the panel was heavily weighted towards instruction, leaving the other two groups with too few Round 2 respondents. Figure 8 shows the distributions.

These are null results from a small and unevenly distributed sample, and their evidential value is correspondingly limited: with 19 raters split across three categories, only a very large association would have been detectable, and the absence of a significant difference is not evidence that no association exists. We report them because they bear directly on the practical question of whether severity can be pre-empted by selection criteria, because the literature on rater effects does not identify stable predictors of severity [9], [10], [57], and because null results of this kind are systematically under reported in Delphi work. On the present evidence, severity behaves as an individual characteristic that panel composition rules do not obviously capture, which is a reason to handle it analytically rather than a demonstration that selection cannot help.

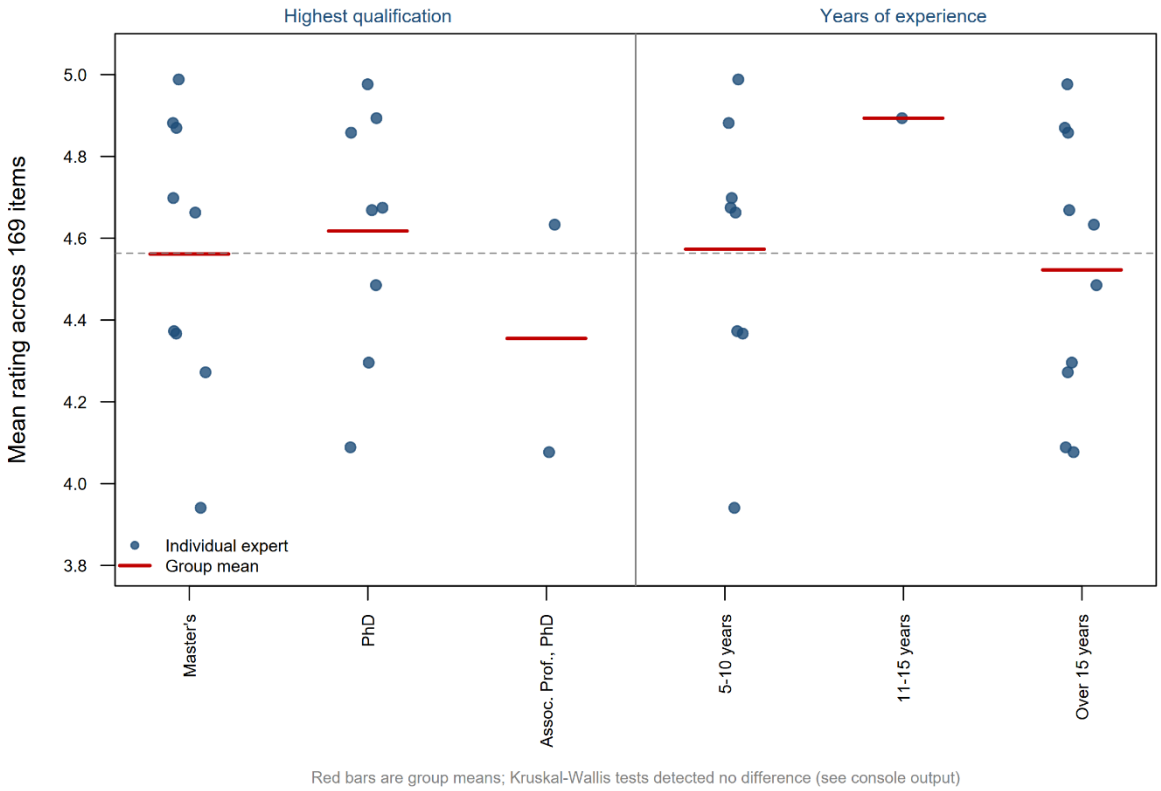


Figure 8. Severity by highest qualification and by years of experience; red bars mark group means.

3.8. What the Findings Do and Do Not Imply

The primary validation conclusion stands, and the two additional specifications reported in Section 3.4 strengthen rather than qualify it. Judged against criteria fixed before data collection and against the guidance those criteria come from [1]–[3], and now also against a chance corrected index and an index that uses the full ordinal range, the framework achieved content validity. What the re-analysis establishes is narrower and, we think, more useful: the headline number is a function of an analytic decision that is almost never reported as such, and the index is nearly silent about the item level differences that researchers most often want to read out of it.

Two of our findings cut against the interpretation an author might prefer, and both should be stated plainly. The panel's agreement about item ordering is real, which removes the possibility that low correlations merely reflect a blunt instrument; and it is negligible in size, which removes the justification for item level ranking. Together these mean that the appropriate use of a content validity index is the global claim it was designed for, namely that a panel of experts finds an instrument acceptable, and not the item level discrimination it is routinely asked to perform.

3.9. Implications For Reporting Practice

We propose that any study reporting a content validity index also report a minimum diagnostic set, none of which requires additional data collection. First, the marginal distribution of all ratings, which reveals ceiling directly and, as Section 3.3 shows, indexes the size of the cut point vulnerability. Second, per rater summaries, namely mean, within rater dispersion and top category use, which make severity visible and flag near constant response styles. Third, the conclusion recomputed under at least one specification that does not dichotomise, for which Aiken's V [12], [13] is the obvious candidate since it requires no additional information and is reported on the same table. Fourth, pairwise rank agreement referred to a permutation null that preserves each rater's marginals, which is the only way to tell whether weak coefficients reflect disagreement or attenuation. A study that reports these four alongside its I-CVI table allows a reader to judge how much of the warrant rests on the analytic convention; a study that reports only the index does not. We note that the modified kappa [2], although frequently recommended, added nothing at this level of agreement, and we would not include it in a minimum set on the strength of these data.

Where item level decisions genuinely must be made, that is, which criteria to reword and which indicators to drop, a measurement model is the appropriate tool, because severity is precisely what such models estimate rather than absorb. Many facet Rasch measurement places raters, items and facets on a common scale and reports item quality adjusted for the severity of whoever happened to rate it [59]–[61], which is the

parameter the content validity index cannot recover; the design of such a model for the present framework is developed in a companion paper [45]. Our findings suggest a division of labour rather than a replacement: an ordinal content validity index remains a reasonable warrant for the global claim that a panel finds an instrument acceptable, while item level refinement requires a model that does not discard the information the refinement depends on.

These recommendations have a direct counterpart in educational assessment policy. The Standards [31] already ask that evidence based on test content describe how experts were selected and how their judgements were elicited; the diagnostic set proposed here extends that expectation from the collection of judgements to the conventions used to summarise them. Three applications follow. First, education journals that publish instrument validation studies could ask authors to report the rating distribution and the per rater summaries in a supplementary table, in the same way that reporting guidance for Delphi studies already asks authors to state how consensus was defined [22], [44]. Second, bodies that approve competence frameworks, learning outcome statements or assessment rubrics for program accreditation or professional certification could treat an index reported without these diagnostics as incomplete evidence, since a framework whose elements the panel cannot distinguish gives weak grounds for deciding which competences a program must assess. Third, institutions building competence based curricula in computing [33], [34] could require the measurement model step described above before a validated framework is converted into graded rubrics [36], [37], because that conversion is the point at which item level differences begin to affect students' marks.

3.10. Limitations

Four limitations bound these claims. First, the findings come from a single dataset in a single domain with 20 raters, so they demonstrate that the vulnerability is real and consequential in a realistic case, not how prevalent it is; replication across datasets with different ceiling levels is the natural next step, and datasets with less ceiling should show a weaker cut point effect. Second, the panel was assembled purposively and drawn from one national system, and its heavy weighting towards instruction limits what the group comparison in Section 3.7 could test. Third, the analysis is a secondary analysis: the instrument was designed for validation, not for studying rater behaviour, so severity is estimated from ratings of items that vary in quality rather than from anchor items of known difficulty, and the permutation model in Section 3.5 tests only whether shared ordering exists, not what the experts were ordering on. Fourth, we treat the five point response as ordinal throughout and report non-parametric tests accordingly, but the facet comparisons and Aiken's V both use means of ordinal ratings, a common practice that nevertheless assumes more than the data strictly license.

4. CONCLUSION

A content validity index of 0.980 and a unanimous consensus across 169 items proved, in this dataset, to be robust against every panel composition test we could apply and fragile against a single change in where an ordinal judgement is dichotomised. Rater severity spanning more than a scale point was fully absorbed by that dichotomisation. Two further specifications located the fragility precisely: a chance corrected index left the conclusion unchanged, and Aiken's V, which never dichotomises, returned a mean of 0.896 with a minimum of 0.788. The instability therefore belongs to the act of dichotomising rather than to the expert judgements themselves. A permutation null model showed that the panel's agreement about item ordering is statistically real yet negligible in magnitude, and one substantive signal, the systematic weakness of measurability relative to relevance, clarity and necessity, survived every specification and is therefore the part of these data on which substantive conclusions can safely rest.

The recommendations follow directly from the three research questions. Report the marginal distribution, the per rater summaries, at least one index that uses the full ordinal range, and pairwise rank agreement referred to a permutation null, alongside every content validity table; and reserve item level refinement for a model that estimates rater severity instead of discarding it. Future research should extend this work in four directions. The first is replication across validation datasets with differing ceiling levels, which would establish whether the share of ratings immediately above the cut predicts the size of the cut point effect, as Section 3.3 implies it should. The second is the application of many facet Rasch measurement to the same matrix, which would allow item quality and rater severity to be estimated jointly and would test whether the weak ordering signal detected here becomes usable once severity is modelled explicitly. The third is instrument design: if the ceiling is what suppresses the ordering signal, then Delphi instruments that force discrimination, through ranking tasks or anchored comparative judgements, may recover at the design stage what no index can recover after the fact. The fourth is specific to education: studies could examine whether frameworks and rubrics validated with the full diagnostic set yield more consistent grading of student work than those validated with the index alone.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the 23 experts from universities who agreed to serve on the Delphi panel and generously devoted their time and expertise to completing two rounds of in-depth evaluation. Their insightful feedback, comments, and suggestions made a significant contribution to refining the study and strengthening its academic rigour. The authors also extend their sincere appreciation to colleagues from higher education institutions who provided valuable input on earlier versions of the manuscript, as well as to the anonymous reviewers for their careful reading, thoughtful evaluation, and constructive comments.

REFERENCES

[1] D. F. Polit and C. T. Beck, "The content validity index: Are you sure you know what's being reported? Critique and recommendations," *Research in Nursing & Health*, vol. 29, no. 5, pp. 489–497, Oct. 2006, doi: 10.1002/nur.20147.

[2] D. F. Polit, C. T. Beck, and S. V. Owen, "Is the CVI an acceptable indicator of content validity? Appraisal and recommendations," *Research in Nursing & Health*, vol. 30, no. 4, pp. 459–467, Aug. 2007, doi: 10.1002/nur.20199.

[3] M. R. Lynn, "Determination and quantification of content validity," *Nursing Research*, vol. 35, no. 6, pp. 382–386, Nov. 1986, doi: 10.1097/00006199-198611000-00017.

[4] E. Almanasreh, R. Moles, and T. F. Chen, "Evaluation of methods used for estimating content validity," *Research in Social and Administrative Pharmacy*, vol. 15, no. 2, pp. 214–221, Feb. 2019, doi: 10.1016/j.sapharm.2018.03.066.

[5] M. S. B. Yusoff, "ABC of content validation and content validity index calculation," *Education in Medicine Journal*, vol. 11, no. 2, pp. 49–54, Jun. 2019, doi: 10.21315/eimj2019.11.2.6.

[6] F. E. Saal, R. G. Downey, and M. A. Lahey, "Rating the ratings: Assessing the psychometric quality of rating data," *Psychological Bulletin*, vol. 88, no. 2, pp. 413–428, Sep. 1980, doi: 10.1037/0033-2909.88.2.413.

[7] C. M. Myford and E. W. Wolfe, "Detecting and measuring rater effects using many-facet Rasch measurement: Part I," *Journal of Applied Measurement*, vol. 4, no. 4, pp. 386–422, 2003.

[8] C. M. Myford and E. W. Wolfe, "Detecting and measuring rater effects using many-facet Rasch measurement: Part II," *Journal of Applied Measurement*, vol. 5, no. 2, pp. 189–227, 2004.

[9] S. A. Wind and M. E. Peterson, "A systematic review of methods for evaluating rating quality in language assessment," *Language Testing*, vol. 35, no. 2, pp. 161–192, Apr. 2018, doi: 10.1177/0265532216686999.

[10] S. A. Wind, "Examining the Impacts of Rater Effects in Performance Assessments," *Applied Psychological Measurement*, vol. 43, no. 2, pp. 159–171, Mar. 2019, doi: 10.1177/0146621618789391.

[11] C. H. Lawshe, "A quantitative approach to content validity," *Personnel Psychology*, vol. 28, no. 4, pp. 563–575, Dec. 1975, doi: 10.1111/j.1744-6570.1975.tb01393.x.

[12] L. R. Aiken, "Content validity and reliability of single items or questionnaires," *Educational and Psychological Measurement*, vol. 40, no. 4, pp. 955–959, Dec. 1980, doi: 10.1177/001316448004000419.

[13] L. R. Aiken, "Three coefficients for analyzing the reliability and validity of ratings," *Educational and Psychological Measurement*, vol. 45, no. 1, pp. 131–142, Mar. 1985, doi: 10.1177/0013164485451012.

[14] V. Zamanzadeh, A. Ghahramanian, M. Rassouli, A. Abbaszadeh, H. Alavi-Majd, and A.-R. Nikanfar, "Design and implementation content validity study: development of an instrument for measuring patient-centered communication," *Journal of Caring Sciences*, vol. 4, no. 2, pp. 165–178, Jun. 2015, doi: 10.15171/jcs.2015.017.

[15] D. M. Rubio, M. Berg-Weger, S. S. Tebb, E. S. Lee, and S. Rauch, "Objectifying content validity: Conducting a content validity study in social work research," *Social Work Research*, vol. 27, no. 2, pp. 94–104, Jun. 2003, doi: 10.1093/swr/27.2.94.

[16] L. L. Davis, "Instrument review: Getting the most from a panel of experts," *Applied Nursing Research*, vol. 5, no. 4, pp. 194–197, Nov. 1992, doi: 10.1016/S0897-1897(05)80008-4.

[17] S. G. Sireci, "Gathering and analyzing content validity data," *Educational Assessment*, vol. 5, no. 4, pp. 299–321, Oct. 1998, doi: 10.1207/s15326977ea0504_2.

[18] S. G. Sireci, "The construct of content validity," *Social Indicators Research*, vol. 45, no. 1–3, pp. 83–117, Nov. 1998, doi: 10.1023/A:1006985528729.

[19] S. N. Haynes, D. C. S. Richard, and E. S. Kubany, "Content validity in psychological assessment: A functional approach to concepts and methods," *Psychological Assessment*, vol. 7, no. 3, pp. 238–247, Sep. 1995, doi: 10.1037/1040-3590.7.3.238.

[20] S. Messick, "Validity," *ETS Research Report Series*, vol. 1987, no. 2, pp. i–208, 1987, doi: 10.1002/j.2330-8516.1987.tb00244.x.

[21] H. A. Von Der Gracht, "Consensus measurement in Delphi studies," *Technological Forecasting and Social Change*, vol. 79, no. 8, pp. 1525–1536, Oct. 2012, doi: 10.1016/j.techfore.2012.04.013.

[22] I. R. Diamond *et al.*, "Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies," *Journal of Clinical Epidemiology*, vol. 67, no. 4, pp. 401–409, Apr. 2014, doi: 10.1016/j.jclinepi.2013.12.002.

[23] R. Boulkedid, H. Abdoul, M. Loustau, O. Sibony, and C. Alberti, "Using and reporting the delphi method for selecting healthcare quality indicators: a systematic review," *PLoS ONE*, vol. 6, no. 6, p. e20476, Jun. 2011, doi: 10.1371/journal.pone.0020476.

[24] P. Nasa, R. Jain, and D. Juneja, "Delphi methodology in healthcare research: How to decide its appropriateness," *World Journal of Methodology*, vol. 11, no. 4, pp. 116–129, Jul. 2021, doi: 10.5662/wjm.v11.i4.116.

[25] M. Niederberger, S. Köberich, and members of the DeWiss Network, "Coming to consensus: the Delphi technique," *European Journal of Cardiovascular Nursing*, vol. 20, no. 7, pp. 692–695, Oct. 2021, doi: 10.1093/eurjcn/zvab059.

[26] J. Spranger, A. Homberg, M. Sonnberger, and M. Niederberger, "Reporting guidelines for Delphi techniques in health sciences: A methodological review," *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, vol. 172, pp. 1–11, Aug. 2022, doi: 10.1016/j.zefq.2022.04.025.
[27] N. Dalkey and O. Helmer, "An experimental application of the delphi method to the use of experts," *Management Science*, vol. 9, no. 3, pp. 458–467, Apr. 1963, doi: 10.1287/mnsc.9.3.458.
[28] C.-C. Hsu and B. A. Sandford, "The delphi technique: making sense of consensus," *Practical Assessment, Research, and Evaluation*, vol. 12, Art. no. 10, 2007, doi: 10.7275/pdz9-th90.
[29] M. T. Kane, "Validating the interpretations and uses of test scores," *Journal of Educational Measurement*, vol. 50, no. 1, pp. 1–73, 2013, doi: 10.1111/jedm.12000.
[30] S. Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning," *American Psychologist*, vol. 50, no. 9, pp. 741–749, 1995, doi: 10.1037/0003-066X.50.9.741.
[31] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*. Washington, DC, USA: American Educational Research Association, 2014.
[32] S. Sireci and M. Faulkner-Bond, "Validity evidence based on test content," *Psicothema*, vol. 26, no. 1, pp. 100–107, 2014, doi: 10.7334/psicothema2013.256.
[33] CC2020 Task Force, *Computing Curricula 2020: Paradigms for Global Computing Education*. New York, NY, USA: ACM, 2020, doi: 10.1145/3467967.
[34] A. N. Kumar *et al.*, *Computer Science Curricula 2023*. New York, NY, USA: ACM, 2024, doi: 10.1145/3664191.
[35] S. Frezza *et al.*, "Modelling competencies for computing education beyond 2020: A research based approach to defining competencies in the computing disciplines," in *Proc. Companion 23rd Annu. ACM Conf. Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion)*, 2018, pp. 148–174, doi: 10.1145/3293881.3295782.
[36] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational Research Review*, vol. 2, no. 2, pp. 130–144, 2007, doi: 10.1016/j.edurev.2007.05.002.
[37] Y. M. Reddy and H. Andrade, "A review of rubric use in higher education," *Assessment & Evaluation in Higher Education*, vol. 35, no. 4, pp. 435–448, 2010, doi: 10.1080/02602930902862859.
[38] M. J. Clayton, "Delphi: A technique to harness expert opinion for critical decision-making tasks in education," *Educational Psychology*, vol. 17, no. 4, pp. 373–386, 1997, doi: 10.1080/0144341970170401.
[39] C. A. Wynd, B. Schmidt, and M. A. Schaefer, "Two quantitative approaches for estimating content validity," *Western Journal of Nursing Research*, vol. 25, no. 5, pp. 508–518, 2003, doi: 10.1177/0193945903252998.
[40] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, "On the practice of dichotomization of quantitative variables," *Psychological Methods*, vol. 7, no. 1, pp. 19–40, 2002, doi: 10.1037/1082-989X.7.1.19.
[41] D. G. Altman and P. Royston, "The cost of dichotomising continuous variables," *BMJ*, vol. 332, no. 7549, p. 1080, 2006, doi: 10.1136/bmj.332.7549.1080.
[42] F. Hasson, S. Keeney, and H. McKenna, "Research guidelines for the Delphi survey technique," *Journal of Advanced Nursing*, vol. 32, no. 4, pp. 1008–1015, 2000, doi: 10.1046/j.1365-2648.2000.t01-1-01567.x.
[43] C. Okoli and S. D. Pawlowski, "The Delphi method as a research tool: An example, design considerations and applications," *Information & Management*, vol. 42, no. 1, pp. 15–29, 2004, doi: 10.1016/j.im.2003.11.002.
[44] S. Jünger, S. A. Payne, J. Brine, L. Radbruch, and S. G. Brearley, "Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review," *Palliative Medicine*, vol. 31, no. 8, pp. 684–706, 2017, doi: 10.1177/0269216317690685.
[45] N. D. Tho, "Designing a SOLO-based analytic rubric for programming projects and a many-facet Rasch framework for its validation," in *International Conference on Educational Measurement and Assessment (VIETAME)*, Vietnam National University Press, to be published 2026.
[46] H. Baumgartner and J.-B. E. M. Steenkamp, "Response styles in marketing research: A cross-national investigation," *Journal of Marketing Research*, vol. 38, no. 2, pp. 143–156, 2001, doi: 10.1509/jmkr.38.2.143.18840.
[47] Y. Van Vaerenbergh and T. D. Thomas, "Response styles in survey research: A literature review of antecedents, consequences, and remedies," *International Journal of Public Opinion Research*, vol. 25, no. 2, pp. 195–217, 2013, doi: 10.1093/ijpor/eds021.
[48] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
[49] R. D. Penfield and P. R. Giacobbi, Jr., "Applying a score confidence interval to Aiken's item content-relevance index," *Measurement in Physical Education and Exercise Science*, vol. 8, no. 4, pp. 213–225, 2004, doi: 10.1207/s15327841mpee0804_3.
[50] P. I. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd ed. New York, NY, USA: Springer, 2005, doi: 10.1007/b138696.
[51] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945, doi: 10.2307/3001968.
[52] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs," *Frontiers in Psychology*, vol. 4, Art. no. 863, 2013, doi: 10.3389/fpsyg.2013.00863.
[53] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937, doi: 10.1080/01621459.1937.10503522.
[54] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961, doi: 10.1080/01621459.1961.10482090.
[55] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952, doi: 10.1080/01621459.1952.10483441.

[56] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2025. [Online]. Available: https://www.R-project.org/
[57] V. Aryadoust, L. Y. Ng, and H. Sayama, "A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research," *Language Testing*, vol. 38, no. 1, pp. 6–40, Jan. 2021, doi: 10.1177/0265532220927487.
[58] G. Engelhard, Jr., "Examining rater errors in the assessment of written composition with a many-faceted Rasch model," *Journal of Educational Measurement*, vol. 31, no. 2, pp. 93–112, 1994, doi: 10.1111/j.1745-3984.1994.tb00436.x.
[59] J. M. Linacre, *Many-Facet Rasch Measurement*. Chicago, IL, USA: MESA Press, 1989.
[60] B. D. Wright and G. N. Masters, *Rating Scale Analysis*. Chicago, IL, USA: MESA Press, 1982.
[61] T. Eckes, *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt am Main, Germany: Peter Lang, 2011, doi: 10.3726/978-3-653-04844-5.

APPENDIX A. THE FRAMEWORK, THE INSTRUMENT AND THE CONSENSUS CRITERIA

This appendix describes the object the expert panel judged and the rules by which its judgements were evaluated. It is included so that the paper can be read and assessed on its own, without reference to any other manuscript. Nothing in it is a result of the present analysis; everything recorded here was fixed before the Round 2 ratings re-analysed above were collected.

A.1. Structure of the framework

The framework describes problem-solving competence in programming for undergraduate information technology students and is organised on three levels. Five competence components describe the phases of the problem-solving process: C1 problem analysis and comprehension, C2 algorithmic solution design, C3 program implementation, C4 testing and debugging, and C5 overall evaluation and optimisation. Twenty-one criteria specify what each component consists of, distributed as four, four, five, four and four criteria respectively. Four developmental levels describe progression within each criterion, and 25 sample behavioural indicators illustrate what each level looks like in observable performance.

The twenty-one criteria are: C1.1 identifying inputs, outputs and constraints; C1.2 decomposing a problem into sub-problems; C1.3 recognising relevant algorithmic patterns and data structures; C1.4 abstracting the problem; C2.1 selecting an appropriate data structure; C2.2 designing a logical and efficient algorithm; C2.3 representing the algorithm in pseudocode or a flowchart; C2.4 analysing and comparing alternative solutions; C3.1 using the syntax of the programming language correctly; C3.2 translating the algorithm into correct source code; C3.3 organising source code clearly and readably; C3.4 using functions and modules effectively; C3.5 reading, understanding and reusing code from libraries and APIs; C4.1 designing test cases; C4.2 detecting and locating errors; C4.3 correcting errors effectively; C4.4 using debugging tools; C5.1 evaluating the correctness of a solution; C5.2 analysing algorithmic efficiency; C5.3 improving and optimising the solution; and C5.4 reflecting and drawing lessons from the process.

The four developmental levels are L1 emerging, where the student requires detailed step-by-step guidance and proceeds by trial and error when no guidance is available; L2 developing, where the student requires algorithmic hints but works from a method rather than by trial and error; L3 proficient, where the student solves independently and formulates questions and locates resources without prompting; and L4 advanced, where the student is fully autonomous and able to guide others.

A.2. The four appraisal facets

Relevance asks whether the element belongs to the construct of problem-solving competence in programming. Clarity asks whether the wording is unambiguous to a reader who did not write it. Necessity asks whether the framework would be incomplete without the element. Measurability asks whether the element can be observed and scored in actual teaching practice. Keeping the facets separate is what makes the facet level comparison in Section 3.6 interpretable, since a panellist could endorse an element as belonging to the construct while doubting that it could be assessed.

A.3. Panel composition

A panel of 23 experts was fixed before Round 1 and drawn purposively from three groups: programming instruction and information technology education in Vietnamese higher education institutions, software industry development and recruitment, and educational measurement and assessment. Inclusion required a postgraduate qualification and at least five years of relevant professional experience. Round 1 returned 21 responses and Round 2 returned 20, of which 19 came from experts who had also responded in Round 1. The panel was weighted towards the first group, which is the limitation noted in Section 3.10 and the reason the professional group comparison in Section 3.7 could not be tested.

A.4. Consensus criteria, fixed in advance

Three criteria were specified before data collection and applied unchanged. First, an item level content validity index of at least 0.78, computed as the proportion of experts rating the item in the upper categories, following the value recommended for panels of more than eight members [1], [3]. Second, ratings of 4 or 5 on the five point scale were treated as indicating that the item satisfied the facet in question. Third, [insert the third criterion exactly as pre-specified]. Every one of the 169 judged items met these criteria. The present paper takes the criteria as given and asks how far the conclusion they produce depends on the second of them, which fixes where an ordinal judgement is collapsed into a binary one.