Building a Formative Assessment Tool Using the CARE² Model: Strategies for Capturing Learner Perceptions of Instructional Design Frameworks

Karen K. Fujii, Niels Brock Copenhagen Business College, Denmark
Natalie Perez, Independent Scholar, United States

The Paris Conference on Education 2026
Official Conference Proceedings

Abstract

Measuring instructional design principles is critical for understanding how pedagogical strategies influence learner experiences, engagement, and outcomes (Spooren et al., 2013). The Fujii and Perez (2025) CARE² model, which includes the principles of Culture, Collaboration, Agency, Agility, Relevance, Repetition, Evaluation, and Engagement, is a culturally responsive framework designed to support international learners in higher education. While the model has guided instructional design, it has not yet been formalized as an evaluative instrument to systematically capture learners' perceptions or experiences. To address this gap, each CARE² principle was operationalized as a formative measurement item, with each dimension representing a distinct, non-interchangeable feature of instructional design rather than a reflective latent trait. Correlation analyses revealed theoretically coherent relationships among constructs (r = .355-.705, p < .001), supporting nomological validity. Importantly, the study also provides a practical template for designers and educators to develop their own evaluative instruments, illustrating how theoretically grounded principles can be translated into measurable indicators. By offering a structured, theory-driven lens, this study supports the evaluation and refinement of pedagogical strategies for designing and enhancing meaningful learning environments, while providing evidence-based guidance for instructional improvement.

Keywords: CARE² model, formative assessment, culturally-responsible pedagogy, learner experience, educational evaluation

Introduction

Course evaluation surveys are commonly used instruments in higher education, playing an important role in instructional evaluation and institutional decision-making (Felder & Brent, 2004; Marsh, 2007; Spooren et al., 2013). Despite their widespread use, the foundations of course measurement are not always examined with the same rigor as other instruments in educational and psychological research (Benton & Cashin, 2012; Uttl et al., 2017). As a result, concerns remain about their validity, including their ability to accurately reflect pedagogical quality and to support fair comparisons across instructors and contexts (Clayson, 2009; Stroebe, 2020). Some argue that these concerns are increased when surveys rely on single-item indicators to capture complex instructional constructs (Bergkvist & Rossiter, 2007; Diamantopoulos et al., 2012). While single-item designs are common in practice due to their efficiency and ease of administration, they raise important methodological questions about reliability and construct representation (Bergkvist & Rossiter, 2007; Diamantopoulos et al., 2012). Traditional psychometric frameworks assume that constructs are best captured through multiple, highly correlated indicators (DeVellis, 2017; Nunnally & Bernstein, 1994). Within this paradigm, single-item measures are often viewed as inherently limited. However, this critique rests on assumptions about the structure of constructs that may not hold in applied educational contexts.

Pedagogical quality is increasingly conceptualized not as a single latent trait, but as a configuration of interrelated instructional design features operating within a learning environment (Doyle, 1986; Lampert, 2001; Wang & Hannafin, 2005). From this perspective, elements such as culture, collaboration, relevance, engagement, and evaluation represent distinct but interacting dimensions rather than interchangeable indicators of one construct, consistent with systems-oriented and design-based approaches to education (Brown, 1992; Cobb et al., 2003; Collins, 1992). For multidimensional constructs, formative measurement models, where indicators define rather than reflect the construct, may be more appropriate (Bollen & Lennox, 1991; Diamantopoulos & Winklhofer, 2001). Misalignment between construct and measurement model can bias estimates and lead to misleading conclusions (Jarvis et al., 2003; MacKenzie et al., 2005), particularly in educational settings where constructs often reflect systems of practice. At the same time, course evaluation research has focused heavily on internal measurement properties while less consistently linking ratings to learning outcomes (Hornstein, 2017; Kane, 2013). Research also suggests that learner evaluations are not always strongly associated with actual learning (Uttl et al., 2017), and comparisons across courses may be confounded by pre-existing learner differences if not properly controlled (Rosenbaum & Rubin, 1983).

This study explores these gaps by developing and validating the CARE² model as a formative, single-item instrument for capturing key dimensions of instructional design. The model includes nine constructs: Culture, Collaboration (peer), Collaboration (instructor), Agility, Agency, Relevance, Repetition, Engagement, and Evaluation, intended to represent core features of the learning environment. Rather than assuming these dimensions reflect a single latent factor, the CARE² model framework conceptualizes pedagogical quality as a composite of distinct but related elements (Fujii & Perez, 2025). To evaluate this approach, the study uses a two-phase validation design that integrates measurement diagnostics with outcome-based survey results. Specifically, it examines whether the CARE² model indicators meet the requirements of formative measurement and whether the resulting composite demonstrates nomological validity by predicting self-reported learning gains (Cronbach & Meehl, 1955).

By linking course evaluation data to external learning outcomes, this study examines the extent to which a course evaluation instrument, using single-items, can be validated or not. The following research questions guide the study:

1. Do the nine CARE² model single-item constructs form a defensible formative composite with acceptable collinearity, outer weights, and sampling adequacy?
2. Does the CARE² model composite demonstrate nomological validity by informing technology learning gains, controlling for pre-existing learner skill levels?

Literature Review

Instructional Measurement in Dynamic Learning Contexts

Educational practice is inherently adaptive, iterative, and context-dependent. Instruction does not function as a fixed intervention but as an evolving system shaped by learners, content, institutional context, and ongoing pedagogical decision-making (Doyle, 1986; Lampert, 2001). This dynamic nature has been widely recognized in design-based research and practitioner inquiry traditions, which emphasize iterative refinement, situated evaluation, and responsiveness to contextual variation as central to educational improvement (Brown, 1992; Cobb et al., 2003; Collins, 1992; Wang & Hannafin, 2005). However, educational measurement has historically been developed under assumptions of stability and generalizability. Most instruments are designed to isolate latent constructs assumed to remain invariant across contexts, enabling psychometric generalization (DeVellis, 2017). This creates a structural mismatch: while instructional systems are adaptive and relational, measurement models are typically static and decomposition-oriented.

Reflective Measurement

Reflective models treat constructs as latent variables that cause observed indicators, which are expected to be interchangeable and highly correlated (Bollen, 1989; DeVellis, 2017). Researchers infer unobservable constructs (e.g., motivation) from observable measures (e.g., survey responses), assuming that changes in the construct are reflected in corresponding changes in indicators (Coltman et al., 2008). Measurement quality is typically assessed via internal consistency (e.g., Cronbach's α), reflecting the assumption of unidimensionality and shared underlying cause (Nunnally & Bernstein, 1994). However, many applied constructs are mis-specified in reflective form when indicators represent distinct conceptual domains rather than interchangeable manifestations (Jarvis et al., 2003). Bollen and Lennox (1991) similarly argue that many constructs are better understood as composites, where indicators define rather than reflect the construct. Reliance on α is further limited by its sensitivity to scale length and tau-equivalence assumptions, and can encourage item redundancy that narrows construct coverage (Cronbach, 1951; Sijtsma, 2009; Streiner, 2003). Consequently, conventional reflective criteria may be poorly suited to multidimensional or system-based constructs.

Formative Measurement

Formative measurement offers a theoretically grounded alternative in which indicators define rather than reflect the construct, with each capturing a distinct component of the phenomenon (Bollen & Lennox, 1991; Diamantopoulos & Winklhofer, 2001). This approach is well suited to composite, multidimensional, and system-oriented constructs, where specification should follow conceptual definition rather than statistical convenience (MacKenzie et al., 2005). When constructs are defined by their components, formative specification is often necessary, and

misalignment between construct and model can lead to invalid inferences even when statistical criteria appear acceptable. Formative measurement also shifts the focus of validity from internal consistency to content validity, ensuring adequate coverage of the construct domain, alongside nomological validity (theoretical relationships) and criterion validity (prediction of outcomes) (Cronbach & Meehl, 1955; Diamantopoulos & Winklhofer, 2001). Together, these criteria provide a more appropriate framework for evaluating measurement quality in complex systems.

Single-Item Measurement

The debate over single-item measurement further illustrates the tension between psychometric rigor and practical utility. Within reflective paradigms, single-item measures are often viewed as inherently inferior due to concerns about reliability and measurement error (Nunnally & Bernstein, 1994). However, this critique assumes that constructs require multiple interchangeable indicators, an assumption that does not hold in formative or diagnostic contexts. Bergkvist and Rossiter (2007) provide empirical evidence that single-item measures can perform comparably to multi-item scales when constructs are concrete and unambiguous. Similarly, Diamantopoulos et al. (2012) argue that the appropriateness of single-item measures depends on the nature of the construct and the purpose of measurement. For evaluative and diagnostic applications, where the goal is to capture overall perceptions or judgments, a single well-designed item may be sufficient.

Methodology

Research Design

This study used a quasi-experimental pre-post design, which is a common approach in educational research when random assignment is not possible (Shadish et al., 2002). A CARE² model course quality survey was administered at the end of the course. In addition, pre-post course technology proficiency surveys were administered at the beginning and end of each course, consistent with standard pretest-posttest approaches for measuring learning gains (Dimitrov & Rumrill, 2003).

Participants

A total of 96 learners completed the CARE² model end-of-course survey, and 91 learners completed both the pre-post course knowledge assessment. Learners were primarily aged 18-24, representing diverse national backgrounds including Nepal and Bangladesh, reflecting the heterogeneity typical of contemporary higher education populations (Canton, 2021).

Instruments

The study used two instruments: the CARE² model survey and a pre- and post-test proficiency survey. The CARE² model survey comprises nine constructs, each measured by a single item on a five-point Likert-type scale (1 = Strongly Disagree, 5 = Strongly Agree). The use of Likert-type scales is standard in educational and psychological measurement for capturing perceptions and attitudes (Edmondson, 2005). The nine constructs include: (1) Culture, (2) Peer Collaboration, (3) Instructor Collaboration, (4) Autonomy, (5) Agility, (6) Relevance, (7) Repetition, (8) Engagement, and (9) Evaluation (see Appendix for instrument details). Each construct is operationalized as a distinct dimension of instructional design, consistent with

formative measurement approaches in which indicators define the construct domain (Bollen & Lennox, 1991; Diamantopoulos & Winklhofer, 2001). The use of single-item indicators aligns with prior research suggesting that single-item measures can be appropriate for concrete, unambiguous constructs, particularly in applied and evaluative contexts (Bergkvist & Rossiter, 2007; Diamantopoulos et al., 2012). In addition, a technology proficiency survey included 11 items assessing digital skills and self-efficacy relevant to the course context. The items used five-point Likert-type response formats. A technology composite score was computed as the mean of all items.

Analytical Strategy

The study examined data across two phases. The first phase examined the CARE² model survey using formative measurement. Several analyses were used to evaluate formative indicators: descriptive statistics, correlation matrices, and multicollinearity diagnostics including variance inflation factor (VIF), tolerance, and condition index (Hair et al., 2018). Outer weights and loadings were examined to assess indicator contributions, with bootstrapping (B = 5,000) used to evaluate statistical significance, consistent with recommended practices in formative measurement and partial least squares modeling (Hair et al., 2018). Harman's single-factor test was also conducted to assess potential common method bias, a standard diagnostic in survey-based research (Podsakoff et al., 2003). Independent samples t-tests were used to compare CARE² model construct scores across the two groups, consistent with standard procedures for group mean comparison (Ary et al., 2018). The second phase examined the nomological validity, which included pretest equivalence testing, within-group pre-post comparisons, and between-group comparisons of post-test scores.

Results

Descriptive Statistics

Descriptive statistics for the CARE² dimensions showed mean scores ranging from 3.72 (SD = 0.97) for Evaluation to 4.08 (SD = 0.74) for Instructor Collaboration, reflecting generally favorable learner perceptions (Table 1). All items spanned the full response range and demonstrated acceptable distributional properties. Skewness (-1.18 to -0.64) and excess kurtosis (0.21 to 2.92) were within recommended thresholds for Likert-type measures, supporting the use of parametric analyses (Field, 2018; Hair et al., 2018). See Table 1.

Table 1
Descriptive Statistics for CARE² Model Single-Item Constructs (N = 96)

Construct	M	SD	Min	Max	Skewness	Kurtosis
Culture	3.80	0.86	1	5	-0.77	1.02
Collaboration_Peer	3.89	0.81	1	5	-1.18	2.76
Collaboration_Instructor	4.08	0.74	1	5	-0.82	1.66
Autonomy	3.95	0.75	1	5	-0.65	1.10
Agility	3.89	0.75	1	5	-1.03	2.92
Relevance	4.06	0.76	1	5	-0.91	1.69

Repetition	3.84	0.87	1	5	-0.64	0.21
Engagement	3.94	0.69	1	5	-0.75	1.96
Evaluation	3.72	0.97	1	5	-0.76	0.32

Pearson Correlation Matrix

Pearson correlations among the nine CARE² model indicators reveal correlations ranging from r = .355 (Culture with Peer Collaboration and Agility) to r = .719 (Instructor Collaboration with Autonomy), with all associations statistically significant ($p < .001$) (see Table 2). Moderate correlations were observed among Instructor Collaboration, Autonomy, Agility, Relevance, and Evaluation (r = .602-.719), suggesting related yet distinct instructor and process-oriented dimensions. Culture showed the weakest associations with other indicators (r = .35-.42). Importantly, no correlation approached the .90 threshold indicative of multicollinearity or redundancy (Hair et al., 2018), supporting the formative assumption that the CARE² model indicators are related but nonredundant. See Table 2.

Table 2

Pearson Correlation Matrix Among CARE² Model Formative Indicators

Construct	C	CP	CI	Au	Ag	Re	Rp	En	Ev
Culture	1.000	0.355	0.413	0.413	0.356	0.395	0.371	0.424	0.406
Collab_Peer		1.000	0.557	0.530	0.478	0.473	0.367	0.470	0.444
Collab_Instr			1.000	0.719	0.625	0.602	0.516	0.528	0.531
Autonomy				1.000	0.642	0.661	0.531	0.589	0.628
Agility					1.000	0.647	0.557	0.608	0.705
Relevance						1.000	0.602	0.620	0.648
Repetition							1.000	0.542	0.576
Engagement								1.000	0.640
Evaluation									1.000

Note. C = Culture; CP = Collab_Peer; CI = Collab_Instructor; Au = Autonomy; Ag = Agility; Re = Relevance; Rp = Repetition; En = Engagement; Ev = Evaluation. Upper triangle only; lower triangle intentionally blank.

Variance Inflation Factor and Tolerance

The findings of the VIF and Tolerance analyses show that all nine VIF values fell below the formative model threshold of 3.3 (Hair et al., 2018), ranging from 1.350 (Culture) to 2.800 (Autonomy). In addition, the tolerance values all exceeded .35 (see Table 3). These results indicate that multicollinearity is not a concern, meaning that each retains sufficient unique variance to contribute a distinct, interpretable signal to the composite. Although Autonomy, Agility, Evaluation, and Instructor Collaboration show relatively higher VIF values (2.51-2.80), this pattern is consistent with the moderate intercorrelations observed among these constructs and remains well within acceptable limits for formative specification. See Table 3.

Table 3
Variance Inflation Factor (VIF) and Tolerance

Construct	VIF	Tolerance
Culture	1.350	0.741
Collaboration_Peer	1.611	0.621
Collaboration_Instructor	2.542	0.393
Autonomy	2.800	0.357
Agility	2.638	0.379
Relevance	2.514	0.398
Repetition	1.854	0.539
Engagement	2.159	0.463
Evaluation	2.623	0.381

Note. VIF threshold for formative models: < 3.3 (2022). Tolerance = 1/VIF.

Condition Index

The eigenvalue-based condition indices independently confirmed the absence of problematic collinearity (see Table 4). The maximum condition index was 4.705 (Dimension 9), which is well below the 15.0 lower warning threshold (Kim, 2019). The first eigenvalue (5.316) captured the dominant shared variance among indicators, consistent with the theoretical expectation of a coherent composite, while the remaining eigenvalues distributed variance across eight additional dimensions, confirming each indicator's unique contribution. See Table 4.

Table 4
Eigenvalues and Condition Indices

Dimension	Eigenvalue	Condition Index
1	5.316	1.000
2	0.747	2.669
3	0.699	2.757
4	0.500	3.260
5	0.467	3.374
6	0.386	3.713
7	0.345	3.924
8	0.300	4.212
9	0.240	4.705

Note. CI = √(λ_max / λ_k). Thresholds: < 15 = no collinearity; 15–30 = moderate; > 30 = severe.

Outer Weights and Bootstrapped Significance

All outer weights were statistically significant (ps < .001), with t-statistics ranging from 8.248 (Culture) to 22.527 (Autonomy), and none of the 95% confidence intervals included zero. Outer loadings ranged from 0.601 (Culture) to 0.830 (Autonomy), exceeding the recommended minimum of .50 (Hair et al., 2018), with seven surpassing .70. Only Culture ($\lambda = 0.601$) and Peer Collaboration ($\lambda = 0.679$) fell in the moderate range. Overall, the results support retaining all nine indicators in the formative composite. See Table 5.

Table 5

Outer Weights, Outer Loadings, and Bootstrapped Significance (B = 5,000)

Construct	Outer Weight	Outer Loading	Bootstrap SE	t	p	95% CI
Culture	0.4593	0.6008	0.095	8.248	< .001	[0.597, 0.974]
Collaboration_Peer	0.5194	0.6794	0.079	11.261	< .001	[0.730, 1.044]
Collaboration_Instructor	0.6101	0.7981	0.063	16.588	< .001	[0.913, 1.163]
Autonomy	0.6348	0.8303	0.048	22.527	< .001	[0.988, 1.178]
Agility	0.6243	0.8166	0.081	13.141	< .001	[0.907, 1.228]
Relevance	0.6276	0.8210	0.057	18.789	< .001	[0.955, 1.181]
Repetition	0.5625	0.7358	0.067	14.311	< .001	[0.829, 1.095]
Engagement	0.6023	0.7878	0.067	15.461	< .001	[0.893, 1.156]
Evaluation	0.6198	0.8107	0.071	14.951	< .001	[0.936, 1.210]

Note. Outer loading = Pearson r(indicator, unit-weighted composite). All p-values < .001 (two-tailed). CI = 95% percentile bootstrap interval.

Common Method Bias: Harman's Single Factor Test

When applying Harman's single-factor test results, the first unrotated factor accounted for 59.07% of total variance across the nine CARE² model items, exceeding the 50% benchmark (Podsakoff et al., 2003). While this may suggest a general factor, these results are expected given that the items reflect related dimensions of a higher-order construct (perceived pedagogical quality) and share a common survey format. Moreover, Harman's test is widely regarded as a limited diagnostic and not definitive evidence of common method bias (Podsakoff et al., 2003). See Table 6.

Table 6

Harman's Single Factor Test: Unrotated Principal Component Solution

Factor	Eigenvalue	% Variance	Cumulative %
F1	5.316	59.07	59.07
F2	0.747	8.30	67.37
F3	0.699	7.77	75.14
F4	0.500	5.56	80.69
F5	0.467	5.19	85.88
F6	0.386	4.28	90.17
F7	0.345	3.84	94.00
F8	0.300	3.33	97.33
F9	0.240	2.67	100.00

Note. CMB threshold: first factor > 50% (Podsakoff et al., 2003).

CARE² Model Construct-Technology Nomological Alignment

Nomological alignment results between the CARE² model constructs and technology survey items show that six of the nine CARE² model constructs map directly onto technology outcomes: Autonomy (digital self-sufficiency), Agility (ease of learning new technology), Relevance (technology awareness and internet research confidence), Repetition (PowerPoint skill acquisition through practice), Engagement (LMS navigation), and Evaluation (online assessment submission). Culture, Peer Collaboration, and Instructor Collaboration are interpersonal dimensions without direct technological counterparts and are therefore represented indirectly within the composite. See Table 7.

Table 7

Nomological Alignment of CARE² Model Constructs With Technology Outcome Items

CARE² Construct	Technology Outcome Item(s)	Theoretical Rationale	Alignment
Culture	N/A	Learning culture is contextual; no direct technology item captures it	Indirect
Collaboration_ Peer	N/A	Peer collaboration not directly measured in technology survey	Indirect
Collaboration_ Instructor	N/A	Instructor collaboration not captured in technology items	Indirect
Autonomy	Sufficient digital skills	Self-sufficiency in digital tools reflects autonomous capability	Direct
Agility	Learn new technologies	Ease of learning new tech directly operationalizes agility	Direct

Relevance	Follow new tech; Internet search confidence	Connecting content to current technology reflects relevance	Direct
Repetition	PPT experience (×4 sub-skills)	Repeated skill practice drives measurable tech proficiency gains	Direct
Engagement	Navigate Moodle; Submit online	Active LMS use and online submission indicate engaged participation	Direct
Evaluation	Submit assessment online	Successful submission is the proximal indicator of assessment completion	Direct

Note. Direct = specific technology item corresponds to the CARE² model construct. Indirect = no direct item; represented through composite.

Pre-Post Technology Change

When examining the pre-post results, paired t-tests for 91 matched learners show that eight of eleven items revealed significant improvement, with the largest gains in online assessment submission ($\Delta = +1.40$, $d = 1.250$), PowerPoint narration ($\Delta = +0.90$, $d = 0.956$), insertion ($\Delta = +1.10$, $d = 0.938$), and saving ($\Delta = +1.19$, $d = 0.969$), indicating strong skill acquisition in course-specific tasks. The overall technology composite increased by 0.631 SDs ($t(90) = -9.274$, $p < .001$, $d = 1.225$), reflecting a large improvement in proficiency. Non-significant gains in broader skills (e.g., learning new technologies, following new technologies, internet search confidence) suggest these were less directly targeted by the curriculum. See Table 8.

Table 8

Pre-Post Paired t-Tests (n = 91 Matched Pairs)

Technology Item	Pre M (SD)	Post M (SD)	Gain	t	p	Cohen's d
Sufficient digital skills	3.10 (0.97)	3.57 (0.82)	+0.47	-3.893	< .001	0.509
Learn new technologies	3.45 (1.12)	3.62 (1.02)	+0.16	-1.224	.224	0.142
Navigate Moodle (LMS)	3.70 (0.92)	4.21 (0.77)	+0.50	-4.700	< .001	0.534
Submit assessment online	2.55 (0.93)	3.96 (0.77)	+1.40	-11.925	< .001	1.250
Follow new technologies	3.33 (0.97)	3.56 (0.90)	+0.23	-1.890	.062	0.241
Confident internet search	3.40 (0.95)	3.62 (0.96)	+0.22	-1.789	.077	0.209
Library database access	2.64 (1.20)	3.40 (1.03)	+0.74	-5.222	< .001	0.606
PPT: Make presentation	1.69 (0.80)	2.72 (1.10)	+1.04	-6.920	< .001	0.900

PPT: Insert images	1.73 (1.01)	2.82 (1.18)	+1.10	-6.049	< .001	0.938
PPT: Narrate presentation	1.33 (0.68)	2.24 (0.93)	+0.90	-6.827	< .001	0.956
PPT: Save as .pptx	1.83 (1.00)	3.02 (0.99)	+1.19	-7.443	< .001	0.969
Technology Composite	-0.22 (0.52)	0.41 (0.59)	+0.63	-9.274	< .001	1.225

Note. n = 91 matched pairs via learner ID. Paired t-test. Gain = Post M - Pre M. Cohen's d = mean difference / SD of difference scores. ns = $p \geq .05$. *** $p < .001$.

Discussion

This study set out to validate the CARE² model course quality instrument through two complementary lenses: formative measurement diagnostics and nomological validity testing against technology learning gains. The results converge to support three principal conclusions: 1/ the CARE² model formative composite is measurement-model defensible, 2/ it is criterion-valid in that the group with higher CARE² model ratings demonstrated substantially greater technology skill gains. More details about the study results are as follows:

Research Question 1: Formative Measurement Validity

The first phase results establish that the CARE² model indicator set meets all quantitative thresholds for a well-specified formative measurement model. The absence of problematic multicollinearity (all VIFs < 3.3; maximum condition index = 4.71) means that outer weight estimates are stable and interpretable, with each indicator's unique contribution to the composite can be reliably estimated and is not distorted by overlap with co-indicators. The universal statistical significance of outer weights (all ps < .001, bootstrapped B = 5,000) confirms that none of the nine indicators is redundant or expendable. Even the weakest contributors, Culture (outer loading = 0.601) and Peer Collaboration (outer loading = 0.679), still exceed the minimum loading threshold and contribute unique variance absent from the other seven indicators.

In addition, we argue that the relatively low loading of Culture is theoretically meaningful rather than psychometrically problematic. Culture reflects institutional and environmental conditions that transcend any single class or instructor; its weaker correlation with the composite defined by the other eight instructor-modifiable dimensions is what we would expect if it captures a distinct, more institutionally situated facet of the learning environment. We see this distinctiveness as an argument for its retention, since without it, the composite would capture only those dimensions of course quality that are directly under instructor control, omitting the broader cultural context in which instruction occurs.

Research Question 2: Nomological Validity and the Meaning of the Composite

The second-phase results provide the most substantively important finding: the group whose learners perceived higher course quality on the CARE² model composite also demonstrated measurably higher technology learning outcomes. The learners, who entered the class with statistically equivalent technology skills (composite d = -0.07), exited with significantly higher proficiency (composite d = 0.38) and showed larger gains on 8 of 11 technology items. This pattern constitutes group-level nomological validity evidence, suggesting that the CARE²

model composite score is not merely a self-contained self-report measure but corresponds to observable differences in learning.

Implications

The results present several implications that should be addressed, especially given the nature of this study and its two-phased analytical approach. The implications are as follows:

Common Method Bias

The results of Harman's single-factor result (59.07%) warrant cautious interpretation. Although the CARE² model items were collected via a single self-report instrument at a single time point, conditions associated with common method bias (e.g., shared method, uniform response format, and positively worded items) are present, and their potential influence cannot be fully excluded. However, the magnitude of the first factor is also consistent with a coherent construct domain, where substantial shared variance is expected among related dimensions of pedagogical quality. Importantly, the learning outcome results provide partial evidence against a purely method-driven explanation, as CARE² model ratings are aligned with differences in technology learning gains across instructional contexts. This pattern is more consistent with substantive variance in perceived pedagogical quality than with common method bias alone.

Measurement Practice

This study demonstrates that a nine-item, single-item-per-construct survey can satisfy formative measurement requirements when rigorously evaluated. Researchers deploying similar instruments should conduct a full diagnostic report, particularly VIF analysis and bootstrapped outer weight significance, rather than assuming that a clean correlation matrix implies measurement validity. The formative framework is the appropriate default when constructs are theoretically non-interchangeable and single items are used, and its diagnostics are less demanding than those of reflective models precisely because they do not require within-construct reliability.

Limitations

Several limitations constrain the interpretation of these findings. For one, the common method bias, flagged by Harman's single-factor test (first factor = 59.07%), is a plausible concern for the CARE² model's self-report data. Although pre-test equivalence testing showed comparable technology skill levels at class entry, unmeasured confounders, like learner motivation, course load, prior academic performance, or demographic characteristics, may account for some of the observed differences. The study design does not support a causal attribution of the differences in gains to pedagogical quality specifically. The study’s generalizability is limited to the institutional and instructional contexts studied. The learner population (primarily aged 18–24, representing South Asian backgrounds) and the specific technology-focused course context may not generalize to other disciplines, institutions, or demographic groups. Finally, the CARE² model single-item design, while meeting formative measurement standards, does not allow for estimation of measurement error at the construct level, meaning that random measurement error is unquantified and potentially consequential for the precision of composite estimates.

Future Research

Future research should examine multi-item indicators for some CARE² model constructs, particularly Culture and Autonomy, where single-item measures may not fully capture conceptual breadth. While single-item formative measurement is defensible, short multi-item versions would enable more nuanced assessment and convergent validation against established instructional design measures. Further work should also test the relationship between CARE² model dimensions and objective outcomes such as academic performance, course completion, and skill transfer, as the present study focused on evaluative judgments rather than downstream behavioral effects. In addition, procedural refinements are recommended to further reduce potential common method bias, including item randomization, incorporation of reverse-coded items, and temporal separation of measures. Adding two to three items per construct would also enable within-construct reliability estimation and support extension toward a reflective and formative hierarchical model.

Conclusion

This study examined whether the CARE² model, constructed from single-item indicators, can be defended as a valid measure of course quality. The findings support this claim on both measurement and substantive grounds. Formative diagnostics indicate that all nine indicators contribute unique, non-redundant variance to the composite, with acceptable multicollinearity, significant outer weights, and adequate sampling properties. Substantively, learners with comparable baseline technology skills exhibited meaningfully different learning gains depending on their CARE² model ratings, providing nomological evidence that the composite reflects real differences in learning outcomes rather than self-report artifacts. Together, these results support the CARE² model as a defensible formative measure of pedagogical quality that corresponds to meaningful differences in learner learning.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that Grammarly, an AI-assisted writing software, was used to proofread and refine the language of the manuscript. The usage was limited to correcting grammatical and spelling errors and rephrasing statements for accuracy and clarity. The author further declares that, apart from Grammarly, no other AI or AI-assisted technologies have been used to generate content in writing the manuscript. The ideas, design, procedures, findings, analyses, and discussion are original and derived from the appropriate and systematic conduct of the research.

References

Ary, D., Jacobs, L. C., Sorensen Irvine, C., & Walker, D. A. (2018). *Introduction to research in education* (10th ed.). Cengage Learning.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: a summary of research and literature*. IDEA Paper No. 50.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*(2), 175–184. https://doi.org/10.1509/jmkr.44.2.175

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

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

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2

Canton, H. (2021). Organization for economic co-operation and development—OECD. In *The Europa Directory of International Organizations 2021* (pp. 677–687). Routledge.

Clayson, D. E. (2009). Student evaluations of teaching: are they related to what students learn? *Journal of Marketing Education, 31*(1), 16–30. https://doi.org/10.1177/027347530832408

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New Directions in Educational Technology* (pp. 15–22). Springer.

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: two applications of formative measurement. *Journal of Business Research, 61*(12), 1250–1262. https://doi.org/10.1016/j.jbusres.2008.01.013

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203–1218. https://doi.org/10.1016/j.jbusres.2008.01.009

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. *Journal of the Academy of Marketing Science, 40*(3), 434–449. https://doi.org/10.1007/s11747-011-0300-3

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: an alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Dimitrov, D. M., & Rumrill, P. D. Jr. (2003). Pretest-posttest designs and measurement of change. *Work, 20*(2), 159–165. https://doi.org/10.3233/WOR-2003-00285

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 392–431). Macmillan.

Edmondson, D. (2005, May). Likert scales: A history. In *Proceedings of the Conference on Historical Analysis and Research in Marketing* (vol. 12, pp. 127–133).

Field, A. P. (2018). *Discovering statistics using IBM SPSS statistics*. (5th ed). Sage, Newbury Park.

Felder, R. M., & Brent, R. (2004). How to evaluate teaching. *Chemical Engineering Education, 38*(3), 200–202.

Fujii, K., & Perez, N. (2025). The CARE² model: action research study exploring multicultural experiences in higher education. *International Journal for Educational Media and Technology, 19*(2). https://www.ijemt.org/index.php/journal/article/view/382

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2018). When to use and how to report the results of PLS-SEM. *European Business Review, 31*(1), 2–24.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification. *Journal of Consumer Research, 30*(2), 199–218. https://doi.org/10.1086/376806

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558–569. https://doi.org/10.1016/j.indmarman.2023.03.010

Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research. *Journal of Applied Psychology, 90*(4), 710–730. https://doi.org/10.1037/0021-9010.90.4.710

Marsh, H. W. (2007). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness. In *The Scholarship of Teaching and Learning in Higher Education: an evidence-based perspective* (pp. 319–383). Springer Netherlands.

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

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Quasi-experiments: interrupted time-series designs. In *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (pp. 171–205). Houghton Mifflin.

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

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Review of Educational Research, 83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: a theoretical and empirical analysis. *Basic and Applied Social Psychology, 42*(4), 276–294. https://doi.org/10.1080/01973533.2020.1756817

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development, 53*(4), 5–23. https://doi.org/10.1007/BF02504682

Appendix

Instruments: CARE² Model Perception Survey Items

Learners completed a nine-item perception survey designed to evaluate their experiences within the pedagogical model. All items were transformed to align with micro-units as compared with class or course. The items were measured using a 5-point Likert scale ranging from 1 = Strongly Disagree to 5 = Strongly Agree.

1. My background and experiences were valued during the micro-units.
2. Working with peers enhanced my learning in the micro-units.
3. Interactions with the micro-units instructor supported my learning.
4. The micro-units provided clear expectations that encouraged me to take responsibility for learning.
5. The instructor adapted the micro-unit content to meet learners' needs.
6. The micro-unit content was meaningful and connected to final course assessments.
7. Important topics, models, or theories for Assessment 1 were reviewed multiple times to reinforce learning.
8. Interactive activities kept me actively involved in learning.
9. I received timely feedback, based on questions and answers, during the micro-unit sessions that helped me improve my knowledge or skills.

Instruments: Technology Survey Items (Pre and Post)

Learners completed a multiple-item survey measured using 5-point agreement scale ranging from Strongly Disagree to Strongly Agree, with item 7 measured using a 5-point confidence scale ranging from Extremely Unconfident to Extremely Confident. Collectively, these items served as indicators of learners' perceived digital confidence across general technology use, learning management system navigation, online assessment submission, technology awareness, information literacy, and academic research skills.

1. I have sufficient skills in digital technology.
2. I can learn new digital technologies without difficulty.
3. I know how to find and navigate the Moodle room.
4. I know how to turn in my assessment online (by uploading it to LearningZone).
5. I follow important new technologies (like ChatGPT).
6. I feel confident in my searches for information online (research, homework, etc.).
7. How confident are you in accessing and using library databases on your laptop?