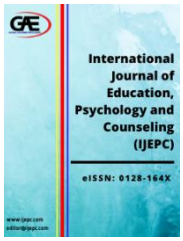
International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090



INTERNATIONAL JOURNAL OF EDUCATION, PSYCHOLOGY AND COUNSELLING (IJEPC)
https://gaexcellence.com/ijepc

DEVELOPMENT AND PRELIMINARY PSYCHOMETRIC EVALUATION OF A CIPP-ALIGNED INSTRUMENT FOR ALTERNATIVE ASSESSMENT IN SPECIAL EDUCATION

Joelene Janes Lim[1]*, Nor Hasnida Che Md Ghazali[2]

[1]Faculty of Human Development, Sultan Idris Education University, Malaysia
p20242002954@siswa.upsi.edu.my https://orcid.org/0009-0001-0538-6462
[2]Faculty of Human Development, Sultan Idris Education University, Malaysia
hasnida@fpm.upsi.edu.my https://orcid.org/0000-0003-1341-9970
*Corresponding Author

Article Info:

Article history:
Received date: 02.08.2026
Revised date: 20.08.2026
Accepted date: 27.09.2026
Published date: 30.09.2026

To cite this document:

Lim, J. J., & Ghazali, N. H. C. M. (2026). Development and Preliminary Psychometric Evaluation of a CIPP-Aligned Instrument for Alternative Assessment in Special Education. *International Journal of Education, Psychology and Counseling, 11*(64), 2071-2090.

Abstract:

Alternative assessment programmes for special educational needs (SEN) pupils have expanded through policy mandates in Malaysia, yet no validated instrument exists for evaluating implementation fidelity across the Context, Input, Process, and Product (CIPP) dimensions. In this study, implementation fidelity refers to the extent to which the Primary Education Alternative Assessment (*Pentaksiran Alternatif Pendidikan Rendah*, PAPR) is enacted according to its intended CIPP components. This study developed and conducted preliminary psychometric evaluation of a CIPP-aligned instrument for assessing PAPR implementation in Special Education Integration Programmes (*Program Pendidikan Khas Integrasi*, PPKI). Item generation drew on literature and policy documents, guided by adult learning theory (informing Input items on teacher readiness and CPD), reflective practice (Process items on daily routines), and competency-based education frameworks (Product items on learning outcomes). Content validity was established through review by six education specialists with 10 to 34 years of experience, yielding item-level Content Validity Index (I-CVI) values of 0.98 to 0.99 and a scale-level S-CVI/Ave of 0.99, exceeding recommended thresholds. A pilot study with 70 special education teachers in Sabah was conducted. Exploratory factor analysis produced a preliminary structure for the instrument with Kaiser-Meyer-Olkin values 0.677 to 0.875, Bartlett's tests were significant ($p < .001$), and the total variance explained ranged from 48.43% to 79.67%. Six items were removed for weak associations with their constructs, leaving 114 items from 120 items that preserved CIPP dimensions coverage. Internal consistency was good to excellent, with Cronbach's alpha values between 0.821 and 0.941. The findings provide initial evidence of content validity and construct structure for a CIPP-aligned PAPR

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

implementation instrument, while confirmatory factor analysis with a larger, stratified sample is recommended before wider deployment.

DOI: 10.35631/IJEPC.1164115

Keyword:

Alternative Assessment, CIPP Model, Instrument Development, Pilot Study, Psychometric Validation, SEN Pupils

© The authors (2026). This is an Open Access article distributed under the terms of the Creative Commons Attribution (CC BY NC) (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact ijepc@gaexcllence.com.

Introduction

Examinations that carry high stakes are losing ground worldwide to assessment that is formative, classroom-embedded and authentic, because such approaches capture what traditional tests cannot see, especially in learners with disabilities and learning difficulties (Black & Wiliam, 2018; Levy-Feldman, 2025). For learners with special educational needs (SEN), the stakes of this change are higher still. Summative tests routinely miss the functional, social and academic progress that best captures how these pupils are actually developing (Quinney, 2021; Ofem & Usoro, 2024). OECD and UNESCO have documented a further obstacle: where infrastructure is weak and technology unevenly available, technology-mediated assessment loses credibility across low- and middle-income countries (OECD, 2020; UNESCO, 2020). Making assessment inclusive is thus an equity question, not only a methodological one, because systems that cannot accommodate diverse learners end up reproducing the very inequalities that inclusive education policies were written to remove (Ang & Phoon, 2025).

Recent assessment reforms have raised the stakes for school-based and alternative assessment programmes for SEN pupils in Malaysia. New alternative assessment pathways introduced while school-based and classroom assessment have been strengthened and selected public examinations have been phased out. These changes rest on policy statements that call for assessment to be continuous, holistic and values-driven, in line with inclusive education and the Malaysia Madani principles of equity and social justice (Ministry of Education Malaysia, 2023, 2026; Ministry of Finance Malaysia, 2023; Ministry of Economy Malaysia, 2023). Within this framework, the Primary Education Alternative Assessment (Pentaksiran Alternatif Pendidikan Rendah, PAPR), replaced the previous Primary School Alternative Assessment (Pentaksiran Alternatif Sekolah Rendah, PASR), to provide a more holistic, competency-based evaluation for SEN pupils in Special Education Integration Programmes (Program Pendidikan Khas Integras*i*, PPKI) schools (Examinations Board, Ministry of Education Malaysia, 2016,

2026). PAPR positions alternative assessment not as a single method but as a cluster of overlapping pathways within the broader School-Based Assessment framework, enacted through observations, performance tasks, portfolios and digital artefacts rather than single summative tests, with expectations that they support ongoing feedback, inclusive classroom participation and credible evidence of achievement (Abdul Manap et al., 2020; Patrick et al., 2026). Evidence from school settings shows that implementing PAPR is influenced persistent rural-urban disparities gaps in connectivity, access to device, ICT support and teacher workload; these factors determine how feasibility and fidelity of implementation is carried out with under resourced schools affected the most (Kasdiah et al., 2024; Penang Institute, 2025). In this study, implementation fidelity refers to how closely PAPR is enacted across the CIPP dimensions from planning, through evidence management, to feedback.

Recent reviews of alternative assessment in teaching practice bring together scattered evidence and reveal a persistent pattern. National policies often promise inclusive, holistic assessment, but classrooms practice is inconsistent. When the gap appears, they are repeatedly explained by the same factors: geographic location, infrastructure provision, and whether teachers receive sustained, assessment-specific preparation versus one-off procedural training (Sanusi et al., 2025; Stufflebeam & Zhang, 2022). These factors align naturally with the CIPP model. Context includes policy mandates and the stubborn digital divide between urban and rural schools; teacher readiness, access to professional development and resource availability make up the input covers teacher readiness, access to professional development, and available resources; processes involve planning, evidence management, and feedback; and products concern perceived fairness, usefulness, and inclusiveness (Abdul Manap et al., 2020; Patrick et al., 2026; Sanusi et al., 2025). The weakest part of this evidence bases most are small-scale, cross-sectional studies that merely measure perceptions rather than working with validated instruments, leaving little that can be said with confidence about how any particular contextual, input or process factor shapes product-level outcomes for SEN pupils (Hair et al., 2019; Ahmad & Ismail, 2020; Sulaiman & Tahar, 2024). The deeper lesson here is that reform without capacity is unlikely to reach classroom: contextual and input conditions decide whether a policy is enacted, and the input-process-product sequence describes by CIPP is not theoretical but observable in schools, enabled or blocked by concrete factors such as digital infrastructure and ongoing professional development (Stufflebeam & Zhang, 2022).

This study therefore aimed to develop and conduct preliminary psychometric evaluation of a CIPP-aligned instrument for assessing PAPR implementation in primary integrated special education schools. The study addresses three research questions: (1) What is the content validity of the CIPP-aligned PAPR instrument as established through expert review? (2) What is the preliminary factor structure identified through exploratory factor analysis? (3) What is the internal consistency reliability of the instrument and its subscales? The study reported here is stage one of a multi-stage development process. It covers item generation from the literature and policy documents, content validation through expert review, and preliminary construct validation through exploratory factor analysis and reliability analysis with a pilot sample of 70 special education teachers in Sabah. Stage two will apply confirmatory factor analysis to a larger stratified random sample of 734 teachers across 134 schools, and stage three will use structural equation modelling and multi-group SEM to test structural relationships and the moderating effect of school location.

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

Literature Review

The CIPP Evaluation Model

Programme quality in the CIPP model is not decided by a single verdict but judged across several interrelated parts, each of domain carries its own evaluation purpose. Context evaluation asks whether the setting and its needs justify the programme; Input evaluation assess the resources, strategies and plans chosen to meet those needs; Process evaluation tracks how PAPR is actually being implemented; and Product evaluation measures outcomes, their sustainability, and transferability to other contexts (Stufflebeam & Zhang, 2022). This comprehensive scope makes the model well suited to educational programmes because it captures both programme results achievement intented and the conditions that enable or obstruct those results (Khamidah & Sanoto, 2025).

Evidence of CIPP's values has grown across the educational, with recent studies using all four dimensions to identify where implementation is strong and where it falters (Agus et al., 2023; Dewi, 2024). Research on inclusive education is revealing by following how policy mandates travel into classrooms, it shows that contextual and input conditions frequently come between a policy intention and its implementation to stall when those conditions fail (Khamidah & Sanoto, 2025). The model strength lies in its analytic logic: it links implementation conditions to outcomes in a decision-oriented way, so an evaluator can explain why a programme succeeded or failed in a specific setting rather than it did (Stufflebeam & Zhang, 2022). Most frameworks stop after measuring the outcomes, CIPP was designed to answer the next question and what caused those outcomes. It offers a theoretically grounded structure for multidimensional assessment programmes by holding the interacting factors of assessment practice across school contexts.

Alternative Assessment for SEN Pupils in Malaysia

Malaysia followed the same direction of global movement, moving its assessment system away from high-stakes examination towards formative, classroom-embedded and authentic practice. Policy has moved from the Primary School Alternative Assessment (Pentaksiran Alternatif Sekolah Rendah, PASR), introduced in 2016, to the Primary Education Alternative Assessment (Pentaksiran Alternatif Pendidikan Rendah, PAPR), a programme revised through successive guidelines and most recently the 2026 edition (Examinations Board, Ministry of Education Malaysia, 2016, 2026). PAPR is positioned as a formal, holistic and learner-centred record of what SEN pupils' abilities across functional, social and academic. Evidence is collected through observations, performance tasks, portfolios and digital artefacts rather than a single summative test, with the expectation that these sources will enable continuous feedback, inclusive classroom participation and reliable records of achievement (Abdul Manap et al., 2020; Patrick et al., 2026). Within the broader School-Based Assessment framework, PAPR operates alongside Classroom-Based Assessment (Pentaksiran Bilik Darjah, PBS), which emphises ongoing formative assessment in inclusive classrooms (Razali et al., 2021). The intended shift is from "assessment of learning" to "assessment for learning", but the literature repeatedly notes tension between the holistic goal and the heavy, rigid documentation burdens that many digital recording platforms place on teachers (Patrick et al., 2026; Sanusi et al., 2025).

The empirical literature shows that implementation challenges match the CIPP dimensions. At the context level, policy mandates run up against persistent rural-urban gaps in digital

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090
provision: many schools still lack reliable internet access, sufficient bandwidth or current ICT equipment (Kasdiah et al., 2024; Penang Institute, 2025). Input factors focus on teacher readiness: where training, resources and administrative support are sufficient, teachers implement more fully; where any of these is missing, readiness for professional development and resource use quickly declines (Abdul Manap et al., 2020). At the process level, teachers generally see PAPR as a positive, holistic innovation, but they struggle with time, weak digital infrastructure and heavy workloads hinder consistent evidence collection and reporting, issues rooted in planning, evidence management and feedback practices (Patrick et al., 2026). Product factors concern perceived fairness, usefulness and inclusiveness, and these perceptions shift according to level support, training and infrastructure available (Sanusi et al., 2025). Altogether, the four CIPP levels reveal a clear pattern: pupils outcome depends on the strong input and process, not on policy mandates alone; the CIPP dimensions stand or fall together.

Instrument Development in Educational Assessment

Instrument development in educational assessment typically follows a clear sequence: items generation from the literature, experts content validation, a pilot test phase with exploratory factor analysis, and a main study using confirmatory factor analysis completes the sequence (Hair et al., 2019; Kline, 2023). Each step prepares the next, so that by the time an instrument is applied it is theoretically grounded, validated by experts for content coverage, and examined psychometrically. Beneath these steps is a basic measurement principle: construct validity is not decided by a single analysis but built from multiple sources of evidence, content evidence, internal structure evidence and response process evidence until the instrument validity is established (Polit & Beck, 2021; Zhao et al., 2025).

Malaysia also offers strong examples of rigorous instrument validation. A quality instrument for teaching, learning and assessment in Malaysian early childhood care and education was validated with polytomous Item Response Theory on a large sample and showed robust psychometric properties (Ahmad et al., 2021). Elsewhere, a scale measuring teachers' feedback conceptions was examined for factor structure and internal consistency, with further analyses deferred until each dimension of feedback belief had empirical support (Ghazali et al., 2019). Taken together, these studies show what measurement rigour benefits when educational constructs are under evaluation: without validated instrument, researchers cannot separate context from process sources and that distinction is exactly what policy makers targeting (Hair et al., 2019; Kline, 2023).

Assessment practice is shaped by interacting the factors, if one-dimensional measure misses that interaction it becomes loss, that's why SEN assessment evaluation requires instrument that cover all four CIPP dimensions. At present, alternative assessment research in SEN settings relies mostly on instruments built for individual studies rather than validated, reusable tools that would allow comparison across studies (Sanusi et al., 2025). This leaves the field unable to build knowledge about what works, for whom and under what conditions. A validated CIPP-aligned instrument would supply the missing common yardstick, letting researchers and educational authorities compare implementation quality across schools, districts and states while also making longitudinal tracking of assessment reform possible (Stufflebeam & Zhang, 2022).

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

Research Gap and Objectives

Limited validated instrument for evaluating PAPR implementation across the four CIPP dimensions in SEN settings could be located, either in the recent systematic review or in the broader literature. Individual studies have examined specific aspects of implementation, including teacher perceptions of inputs and processes (Abdul Manap et al., 2020), evidence management practices (Patrick et al., 2026), and product-level evaluations of fairness and usefulness (Sanusi et al., 2025), but the evidence base remains dominated by small-scale, cross-sectional designs built on perceptions, with limited use of validated instruments (Hair et al., 2019; Ahmad & Ismail, 2020; Sulaiman & Tahar, 2024). Where instruments were properly validated at all, this tended to happen in only a handful of studies, notably Ahmad et al. (2021), who validated theirs through polytomous IRT, and Ghazali et al. (2019), who did so through factor analysis. These cases confirm that rigorous instrument development is both achievable and worthwhile in Malaysian educational research, although PAPR implementation in SEN settings was not the concern of either study.

Methodology

Research Design

This study used a quantitative, cross-sectional survey design to examine the preliminary psychometric properties of a CIPP-aligned instrument for evaluating PAPR implementation in PPKI settings. Cross-sectional surveys support rigorous statistical tests and subgroup comparisons, producing the kind of evidence needed for instrument validation in educational research (Creswell & Creswell, 2023). They also give a practical and robust snapshots of implementation as it currently stands, consistent with standard survey methodology in education.

The Research Onion guided methodological choices in this study, moving from philosophical stance to analytical technique to keep the design coherent, defensible and relevant to policy (Saunders et al., 2019). Within this framework, the study takes a positivist stance that values objectivity, measurement and generalizability qualities appropriate to evaluate programme implementation in school systems (Gray, 2021). The approach is deductive: the instrument structures are drawn from the theoretical CIPP model and then tested with empirical data (Creswell & Creswell, 2023). A mono-method design using a single structured questionnaire keeps measurement and produced data suitable for statistical analysis (Cohen et al., 2018).

This study represents stage one of a multi-stage instrument development sequence. The three research questions produce complementary evidence strands. RQ1 addresses content validity, checking whether items adequately represent the CIPP constructs and the PAPR framework (Polit & Beck, 2021; Zhao et al., 2025). RQ2 examines the preliminary factor structure to identify the underlying measurement model and prepare for later confirmatory analysis (Hair et al., 2019). RQ3 assesses internal consistency reliability to determine whether items within each subscale form coherent constructs (Kline, 2023). Together, these three strands set up stage two, a confirmatory factor analysis using a larger stratified random sample of 734 teachers across 134 schools, and then stage three, which will apply structural equation modelling and multi-group SEM to test the structural relationships and the moderating effect of school location.

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

Participants and Sampling

This study targeted special education teachers in Sabah because they are the one who turn assessment policy into daily classroom practice, and their perspective are the core judgement for the instrument's content validity and preliminary factor structure (Ramli et al., 2023). Eligible special education teachers who had either completed formal PAPR-related training or taken part in the PAPR programme during the preceding twelve months, a condition that kept respondents experience current and relevant (Ramli et al., 2023). The sampling frame covers 134 primary integrated special education schools across the 24 districts of Sabah, together employing 920 teachers. The main study will select 734 special education teachers through stratified random sampling with proportional allocation, ensuring that every district, and both rural and urban locations (Memon et al., 2020; Creswell & Creswell, 2023). The frame was compiled school by school and district by district so that claims of of representativeness rest on documented coverage rather than assumption.

This study represents a pilot test. 150 questionnaires were distributed, 70 were returned and analysed, giving a final sample of 70 special education teachers. This figure falls within the recommended range of 50 to 100 cases for preliminary psychometric work on multidimensional instruments (Hair et al., 2019; Kline, 2023; Kalkbrenner, 2021). The sample is sufficient for exploratory factor analysis and reliability analysis in advance of the main survey. Participants were recruited via a state-level PAPR course for chief and area assessors from across Sabah, professional networks, and approved school visits.

Instrument Development

The instrument was developed by aligning the four CIPP dimensions onto constructs taken from the literature on alternative assessment, inclusive education, and Sabah's special education context (Stufflebeam & Zhang, 2022; Zhao et al., 2025). Draft items came from three sources: relevant policy documents, existing instruments, and empirical studies of PASR, PAPR and classroom-based assessment. Grounding the instrument both in policy and research aimed to ensure methodological rigor, policy relevance and consistency with current evaluation practice in education (Stufflebeam & Zhang, 2022). The tool uses closed Likert-scale items to capture perceptions of implementation fidelity, effectiveness and outcomes, including a brief demographic section to contextualise responses.

Item generation followed a two-pronged approach: potential items were produced inductively from recent empirical studies and then checked deductively against Malaysian education policy documents. Which constructs and indicators were chosen for each CIPP dimension was guided by theory, with policy-implementation and collaborative-governance theories informing Context, adult learning theory informing Input, reflective practice and constructivism informing Process, and competency-based education and inclusive-assessment theories informing Product (Stufflebeam & Zhang, 2022; Knowles et al., 2020; Black & Wiliam, 2022; Mulder, 2023). The initial instrument comprised 120 substantive items organised into four CIPP-based domains: Context (20 items), Input (30 items), Process (35 items), and Product (35 items). All substantive items used a five-point Likert-type scale anchored from 1 (strongly disagree) to 5 (strongly agree), in line with recommendations that multidimensional survey instruments balance construct coverage against respondent burden (Hair et al., 2019; Kline,

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

2023; Tabachnick & Fidell, 2019). Table 1 summarises how each sub-construct was mapped deductively to its governing theory across the four CIPP domains.

Table 1: Item Generation Mapping Across CIPP Domains

Domain	Sub-construct	Items	Item Codes	Theoretical Sources
Context	FPK in PAPR	7	A1-A7	Black & Wiliam (2022); Ang & Phoon (2025); MOE (2023, 2026)
	PAPR Policies	7	A8-A14	Black & Wiliam (2022); Ang & Phoon (2025); MOE (2023, 2026); Stufflebeam & Zhang (2022)
	PAPR Objective	6	A15-A20	Black & Wiliam (2022); Ang & Phoon (2025); MOE (2023, 2026)
Input	Resources	7	B1-B7	Knowles et al. (2020); MOE (2023); Sanusi et al. (2025)
	Teacher Readiness	7	B8-B14	Patrick et al. (2026)
	Stakeholder Readiness	6	B15-B20	Abdul Manap et al. (2020)
	PAPR CPD Hours	10	B21-B30	Mulder (2019)
Process	PAPR Fidelity	12	C1-C12	Stufflebeam & Zhang (2022); Black & Wiliam (2022)
	Teacher's Daily Routines	12	C13-C24	Razali et al. (2021)
	Leaders Monitoring	11	C25-C35	Sanusi et al. (2025)
Product	Student Learning Outcomes	12	D1-D12	Mulder (2019); Ang & Phoon (2025); UNESCO (2020); OECD (2020)
	Inclusive Effectiveness	11	D13-D23	Abdul Manap et al. (2020); Patrick et al. (2026)
	Program Impact	12	D24-D35	Sanusi et al. (2025); MOE (2026)

Source: Field data (2026)

Content Validation

Content validity was examined by a panel of six education specialists made up of PAPR trainers, assessment and examinations officers, and senior officers whose combined experience in assessment and special-education practice ranged from 10 to 34 years. Panel members were chosen purposively using three criteria: direct involvement in PAPR implementation or training; current or past service in a district or state level assessment or examinations unit; and prior experience in developing or evaluating research instruments. The six experts collectively brought policy-level assessment knowledge, practical PAPR implementation experience, and the technical skills required to review a measurement tool (Ying et al., 2025). Table 2 provides the expert panel profiles.

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

Table 2: Expert Validation Panel Composition

No.	Expert/Officer	Position	Years of Experience	Aspects Evaluated
1	En Mohammad Din Bin Mohd Yassin	Head Sector of Assessment and Examinations Unit, Keningau District Education Office	34	Content relevance
2	En Stephen Bin Ansim	Assistant Officer for Assessment and Examinations Unit, Keningau District Education Office	32	Content relevance
3	Dr Osman Bin Said	Assistant Officer for Special Education Learning Unit, Keningau District Education Office	31	Content relevance
4	En Amirul Farhan Bin Ahmad Murad	Chief Assessor for PAPR, Keningau District	11	Content relevance
5	En Mohd Sahabudin Bin Ayub	Assistant Assessor for PAPR, Keningau District	10	Content relevance
6	Pn Saleha Binti Mersin @Wainon	School Improvement Specialist Coach (SISC+), Keningau District Education Office	28	Language clarity

Source: Field data (2026)

The Content Validity Index (CVI) and the Content Validity Ratio (CVR) were computed for each item and for the overall scale from the six experts' ratings on a three-point scale (1 = not suitable, 2 = moderate, 3 = suitable), quantifying agreement on content relevance and language clarity (Lawshe, 1975), and an item was kept only if at least five of the six panel members rated it essential (Lawshe, 1975).

Pilot Study Procedure

A total of 150 pilot questionnaires were distributed, and 70 completed forms were returned. Respondents were special education teachers recruited in three ways: at state-level PAPR course for chief and area assessors held across Sabah, via professional networks, and through school visits conducted with prior approval from headmasters. Questionnaires were handed out and collected during course sessions and professional-network meetings so respondents could ask questions immediately, non-response was minimised, and returned forms were checked for completeness before collection ended. The response rate of about 47% is typical for field-based pilot work, where school timetables and workload routine often limit participation (Kalkbrenner, 2021). The pilot data were then examined for item clarity, distributional properties and preliminary factor structure, consistent with the purpose of pilot studies to refine an instrument before larger scale deployment (Creswell & Creswell, 2023; Hair et al., 2019).

Participation was entirely voluntary. Before making any decision, each respondent received an information sheet explaining the study's purpose, what participation involved, possible risks and benefits, and their rights, including the right to decline or withdraw at any time without penalty. Those who agreed to take part signed written informed consent prior to complete the

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

questionnaire. Confidentiality and anonymity were maintained throughout no personally identifiable information was collected, and school identifiers were coded so that neither schools nor individual teachers could be identified in any reports (MCRCR, 2021). Data collection dates were scheduled in consultation with headmasters to avoid peak teaching and reporting periods, and the burden on participants was kept to a minimum.

Data Analysis

Data analysis ran in three stages for this pilot study, one per research questions. For RQ1, content validity was assessed using the Content Validity Index (I-CVI and S-CVI/Ave) together with the Content Validity Ratio (CVR), both computed from the ratings given by the six experts on a three-point scale (Lawshe, 1975; Polit & Beck, 2021). For RQ2, exploratory factor analysis was performed using principal axis factoring with oblique rotation in line with accepted practice for multidimensional instruments in educational research (Hair et al., 2019; Kline, 2023). EFA suitability was evaluated using the Kaiser-Meyer-Olkin (KMO) index, Bartlett's test of sphericity, items with factor loadings and a total variance explained. For RQ3, internal consistency reliability was assessed using Cronbach's alpha. Normality was assessed through skewness and kurtosis values. All analyses were conducted using IBM SPSS Statistics v30.

Results

Content Validity Results

Content validity was established through expert review by six education specialists with 10 to 34 years of experience in assessment, examinations, and special-education practice. Each expert rated the 120 initial items on a three-point scale (1 = not suitable, 2 = moderate, 3 = suitable), and the Content Validity Index (CVI) and Content Validity Ratio (CVR) were computed for each CIPP dimension and for the overall scale. Table 3 presents the CVI and CVR indices across the four CIPP dimensions.

Table 3: Content Validity Indices Across CIPP Dimensions

Dimension	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Sum of 'Suitable'	I-CVI
Context	20	19	20	20	20	20	119	0.99
Input	30	29	29	29	30	30	177	0.98
Process	35	35	34	35	35	35	209	0.99
Product	35	35	34	35	35	35	209	0.99
Total / Scale-level	120	118	117	119	120	120	714/720	0.99

Source: Field data (2026)

Table 3 shows the item-level CVI (I-CVI) values for the four dimensions ranged from 0.98 to 0.99, exceeding the commonly recommended minimum of 0.78 and indicating excellent item-level content validity (Zhao et al., 2025; Liu et al., 2023). The average scale-level CVI (S-CVI/Ave) of 0.99 surpassed the 0.90 benchmark, demonstrating excellent overall scale-level validity (Zhao et al., 2025; Liu et al., 2023). The CVR value of 0.99 exceeded the minimum threshold of 0.99 for six experts, confirming that only items rated as essential by at least five

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

of six panel members were retained (Lawshe, 1975). These results confirm that the instrument exhibits excellent content validity and is suitable for preliminary evaluation in applied educational settings (Polit & Beck, 2021).

Data Quality Screening

The pilot dataset was screened for missing values, distributional properties and suitability for parametric analysis before exploratory factor analysis. No missing values turned up on any of the key variables, so neither imputation nor casewise deletion was called for, and the analysis kept its full statistical power with little risk of biased estimates further down the line (Hair et al., 2019). Table 4 presents the descriptive statistics and normality indices for each CIPP construct.

Table 4: Descriptive Statistics and Normality Indices

Variable	Mean	Std. Deviation	Skewness	Kurtosis
Context	4.243	0.397	-0.719	1.440
Input	4.174	0.441	-0.423	0.168
Process	4.142	0.429	-0.193	0.270
Product	4.194	0.457	-0.344	0.043

Source: Field data (2026)

Table 4 shows mean scores were high for every construct, between 4.14 and 4.24, well clear of the midpoint and therefore indicating strong agreement by common interpretation (Koo & Yang, 2025). Standard deviations ranged from 0.397 to 0.457, showing the responses were tightly clustered and consistent across the sample (Hair et al., 2019). Skewness values fell between -1 and +1 and kurtosis values between -2 and +2, keeping all distributions within commonly accepted bounds for approximate normality (Kline, 2023). Altogether, these results justify using parametric techniques such as factor analysis, and the tolerance of Likert-scale data for these methods even when parametric assumptions are moderately violated (de Winter & Dodou, 2010).

Exploratory Factor Analysis Results

Exploratory factor analysis (EFA) was conducted to examine the preliminary factor structure of the instrument and to flag items falling short of the psychometric thresholds set in advance. Six items (A7, A10, A20, B4, B15, B16) failed to meet those thresholds and were removed, leaving 114 items with theoretical coverage intact across all four CIPP dimensions. Table 5 presents the EFA and reliability findings by construct and sub-construct.

Table 5: EFA and Reliability Findings by Construct and Sub-construct

Construct	Sub-construct	KMO	Bartlett's Test	TVE	Items Before	Items After	Cronbach's Alpha
Context	FPK in PAPR	0.875	$p < .001$	58.17%	7	6	0.906
	PAPR Policies	0.681	$p < .001$	72.02%	7	6	0.833

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

	PAPR Objective	0.677	p < .001	77.53%	6	5	0.821
Input	Resources	0.818	p < .001	60.50%	7	6	0.912
	Teacher Readiness	0.785	p < .001	61.66%	7	7	0.915
	Stakeholder Readiness	0.759	p < .001	48.43%	6	4	0.839
	PAPR CPD Hours	0.819	p < .001	77.32%	10	10	0.941
Process	PAPR Fidelity	0.751	p < .001	66.19%	12	12	0.907
	Teacher's Daily Routines	0.703	p < .001	67.80%	12	12	0.910
	Leaders Monitoring	0.736	p < .001	79.67%	11	11	0.870
Product	Student Learning Outcomes	0.745	p < .001	68.03%	12	12	0.912
	Inclusive Effectiveness	0.778	p < .001	69.76%	11	11	0.911
	Program Impact	0.788	p < .001	76.83%	12	12	0.929
Total					120	114	0.985

Source: Field data (2026)

Table 5 reports KMO values between 0.677 and 0.875, above 0.60 minimum and within the band read as adequate to excellent for factor analysis (Hair et al., 2019). Bartlett's test of sphericity was significant at the .001 level for every subscale, indicating that the correlation matrices were suitable for extraction. Total variance explained ranged from 48.43% to 79.67%, and while most subscales passed the commonly recommended 50% to 60% threshold, the results overall point to satisfactory construct validity (Hair et al., 2019; Howard, 2023). Table 6 sets out the items removed after pilot testing, their statistical characteristics, and the reasons for removal.

Table 6: Items Deleted After Pilot Testing

Dimension	Construct	Item	Statement	Loading	Communality	Reason for Deletion
Context	A. FPK in PAPR	A7	I find FPK in PAPR strengthens inclusive practices	0.42	0.495	Low communality (< .50), weak association with construct
	B. PAPR Policies	A10	I find FPK in PAPR strengthens	0.32	0.372	Low communality (< .50), weak

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

			inclusive practices			measurement focus
	C. PAPR Objective	A20	I see PAPR matches inclusive education priorities	0.44	0.455	Low communality (< .50), insufficient construct representation
Input	A. Resources	B4	I use school digital infrastructure for PAPR	0.35	0.473	Low communality (< .50), poor construct clarity
	B. Stakeholder Readiness	B15	I see stakeholder collaboration supporting PAPR	0.38	0.248	Low communality (< .50), compromising internal consistency
	B. Stakeholder Readiness	B16	I am aware that parents are regularly updated on PAPR	0.37	0.451	Low communality (< .50), compromising internal consistency

Source: Field data (2026)

Table 6 shows, all six deleted items had factor loadings below 0.50 and communalities below 0.50, pointing to weak associations with their intended constructs and insufficient measurement focus (Hair et al., 2019; Kline, 2023). The instrument's psychometric quality improved after the six items removed with no CIPP dimension or sub-construct lost. The decisions follow the same logic reported in other Malaysian validation studies, where items were retained or removed on statistical grounds and the underlying constructs were then checked for integrity through strong KMO values, substantial explained variance and high reliability (Ahmad et al., 2021; Ghazali et al., 2019).

Reliability Analysis Results

Cronbach's alpha assessed the internal consistency reliability for each CIPP dimension and sub-construct. Table 7 presents the reliability coefficients.

Table 7: Reliability Coefficients by CIPP Dimension and Sub-construct

Dimension	Sub-construct	Cronbach's Alpha	Interpretation
Context	FPK in PAPR	0.906	Excellent
	PAPR Policies	0.833	Good
	PAPR Objective	0.821	Good
	Overall	0.920	Excellent

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

Input	Resources	0.912	Excellent
	Teacher Readiness	0.915	Excellent
	Stakeholder Readiness	0.839	Good
	PAPR CPD Hours	0.941	Excellent
	Overall	0.929	Excellent
Process	PAPR Fidelity	0.907	Excellent
	Teacher's Daily Routines	0.910	Excellent
	Leaders Monitoring	0.870	Good
	Overall	0.955	Excellent
Product	Student Learning Outcomes	0.912	Excellent
	Inclusive Effectiveness	0.911	Excellent
	Program Impact	0.929	Excellent
	Overall	0.962	Excellent
Overall Instrument		**0.985**	**Excellent**

Source: Field data (2026)

Table 7 shows that Cronbach's alpha values at the sub-construct level ranged most above 0.90 for excellent reliability (Hair et al., 2019). The overall instrument alpha of 0.985 supports excellent internal consistency for group-level analysis (Hair et al., 2019; Bujang et al., 2024). The trimmed 114-item scale eased respondent burden while holding measurement precision, which matters in school settings where decisions need to be made on time.

Discussion

Content Validity

The content validity results show excellent expert consensus across all four CIPP dimensions. Item-level CVI (I-CVI) values of 0.98 to 0.99 and a scale-level CVI (S-CVI/Ave) of 0.99 exceed the commonly recommended benchmarks of 0.78 and 0.90, respectively (Polit & Beck, 2021; Zhao et al., 2025). Two features of the development process help explain the strength of this agreement. The first is the dual-source item generation strategy. Items were derived inductively from empirical literature and mapped deductively against Malaysian education policy documents, so each one was anchored in both theoretical constructs and policy requirements (Stufflebeam & Zhang, 2022). The second is the panel itself, PAPR chief assessors, assessment and examinations officers, special education learning unit staff and a School Improvement Specialist Coach with 10 to 34 years of experience, combined policy-level oversight with hands-on implementation knowledge (Ying et al., 2025).

Preliminary Factor Structure

Six items deleted from the exploratory factor analysis and left 114 items without losing any of the four CIPP dimensions or their 13 sub-constructs. The six removals all fell short of the thresholds set in advance, with factor loadings and communalities below 0.50, which is weak association with their intended constructs (Hair et al., 2019; Kline, 2023). KMO values ranged from 0.677 to 0.875, comfortably above the 0.60 minimum, and Bartlett's tests were significant ($p < .001$) for every subscale, so the data were suitable for factor extraction (Hair et al., 2019). Total variance explained ranged from 48.43% to 79.67%; most subscales cleared the 50% to

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

60% level commonly recommended for social-science instruments (Hair et al., 2019; Howard, 2023).

Every item removed had a clear theoretical justification rather than being dropped for statisticl convenience. The three Context items deleted A7, A10 and A20, related to FPK alignment and the PAPR objective, exactly where expert review had earlier identified as conceptually overlapping with other items. Among the Input items, B4 and B15 concerned digital infrastructure and stakeholder collaboration, these constructs will require more context-specific operationalization in any future version. Item B16, which addressed parent awareness, failed to load cleanly on the stakeholder readiness factor in the pilot data. Such deletions mirror common practice reported in Malaysian validation studies, where items with weak communalities or ambiguous loadings are removed to improve factorial clarity and internal consistency (Ahmad et al., 2021; Ghazali et al., 2019). The remaining 114 items displayed acceptable communalities and strong corrected item-total correlations, supporting their retention as indicators of the CIPP constructs they are intended to measure.

Reliability Evidence

Internal consistency was satisfactory for the full instrument and for each subscale, ranging from good to excellent. At sub-construct level, Cronbach's alpha ranged from 0.821 (PAPR Objective) to 0.941 (PAPR CPD Hours); all values exceeded the 0.80 threshold for good reliability, and most exceeded 0.90, a level commonly interpreted as excellent (Hair et al., 2019). The four higher order dimensions produced even stronger alphas: Context 0.920, Input 0.929, Process 0.955 and Product 0.962. Given these results, the pilot fulfilled its purpose, and the logical next steps are confirmatory factor analysis (CFA) and structural equation modelling (SEM) on the main-study sample to refine the instrument further and strengthen its theoretical basis (Hatem et al., 2022; Hair et al., 2019).

Two points merits emphasis. First, within each subscale the items hold together as a single measure, and it is coherence that will let the CIPP dimensions operate as latent in forthcoming CFA and SEM analyses (Kline, 2023). Second, the overall alpha of 0.985, requires more careful interpretation: alpha increases with the number of item count, and with 114 items part of the high value reflects scale length. Even after accounting for that effect, the instrument demonstrates strong consistency across respondents, and a quality programme evaluation needs when its findings are intended to inform policy. The 114-item version achieves this reliability at a reasonable cost in respondent time.

Instrument Contribution

What this study adds is an instrument that is at once CIPP-aligned, SEN-focused, validated through expert review and exploratory factor analysis, and purpose-built for evaluating PAPR implementation. Limited earlier work combines all four features. The closest neighbours each fall short on one of them. Ahmad et al. (2021) took an ECCE quality instrument through polytomous IRT with a large sample, rigorous by any standard, but worked with early childhood populations rather than SEN, and their construct was teaching quality rather than programme evaluation. Ghazali et al. (2019) factor-validated a scale on conceptions of assessment, valuable work for the assessment field but one that never reaches CIPP dimensions or SEN contexts. Abdul Manap et al. (2020) used CIPP as a survey framework without developing a reusable instrument from it, Sanusi et al. (2025) mapped implementation

challenges in a systematic review without converting them into a measurement tool, and Patrick et al. (2026) offered qualitative insight into evidence management without producing a quantitative instrument. Sulaiman and Tahar (2024) studied UDL implementation, with no CIPP alignment and no SEN focus.

Placed alongside related work, the measurement gap is evident, and this instrument addresses that gap rather than duplicating existing tools. It offers a CIPP-aligned, reusable means of validating PAPR implementation in SEN settings and provides a common yardstick for comparing implementation quality across schools, districts and states (Stufflebeam & Zhang, 2022). Two design decisions explain why it captures what other instruments miss: first, its explicit grounding in the CIPP model; second, its dual-source item generation, drawing on empirical studies and Malaysian policy documents so that context, input, process and product factors specific to PAPR implementation are measured directly rather than inferred. The pilot ends this stage of work appropriately: moving to confirmatory factor analysis and structural equation modelling shifts the question from whether the tool describes implementation quality to whether it can explain it, by testing causal pathways and moderating effects (Hair et al., 2019; Kline, 2023).

Limitations

The evidence presented here is preliminary and should be interpreted accordingly, since this study is stage one of a longer instrument development process. Four limitations should be acknowledged. The pilot sampling is the first constraint. Participants were drawn from teachers attending a state-level PAPR course and from professional networks in Sabah, which narrows the study's geographic and contextual reach. The variability suffices for exploratory factor analysis, but not every of Sabah's 24 districts is represented and most notably remote rural and interior schools whose infrastructure exerts a distinct influence on assessment practice (Kasdiah et al., 2024; Penang Institute, 2025). Sample size is the second limitation. 70 special education teachers sit comfortably within the 50–100 range recommended for preliminary psychometric work (Kalkbrenner, 2021; Hair et al., 2019), so the exploratory results have reasonable footing. However, definitive construct validation requires larger samples: confirmatory factor analysis yields stable parameter estimates and reliable fit indices only with greater numbers. The planned main study (n = 734) exceeds common SEM minima and addresses this shortfall (Gulturk, 2024; Hair et al., 2019; Kline, 2023). The third limitation is the cross-sectional design. A single timepoint snapshot cannot assess temporal stability, test and retest reliability, or causal relations among the CIPP dimensions (Hair et al., 2019). The fourth is that the factor structure identified through EFA remains provisional until CFA confirms it. Extracted factors may not replicate exactly in confirmatory analysis, and items may need re-specification or removal based on modification indices and standardised loadings (Kline, 2023; Hair et al., 2019). Lastly, the uniformly positive responses restricted score variability, which may have attenuated item correlations; the strength of the factor solution despite this restriction warrants cautious interpretation.

Conclusion

This study represents stage one of a multi-stage instrument development process and conducted preliminary psychometric evidence from a pilot sample of 70 special education teachers in Sabah evaluating the implementation of PAPR using the CIPP model. Findings are not generalisable nationally until stage two CFA with the stratified sample of 734 teachers. The

CFA in stage two will test whether the factor structure identified through EFA holds in confirmatory analysis, with model fit evaluated on multiple indices including CFI, TLI, RMSEA, and SRMR (Hair et al., 2019; Kline, 2023; Cheung et al., 2024). Stage three will apply structural equation modelling to the structural relationships among the CIPP dimensions, with multi-group SEM examining whether school location (rural versus urban) moderates the input-process-product pathways the CIPP model theorises. Taken together, the validation evidence from stages two and three will determine whether the instrument can move beyond describing implementation quality to explaining it, through tests of causal pathways and the targeting of policy intervention.

Acknowledgements:	The authors would like to express their sincere gratitude to Sultan Idris Education University for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to supervisor Nor Hasnida Che Md Ghazali contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.
Funding Statement:	No Funding
Conflict of Interest Statement:	The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the International Journal of Education, Psychology and Counselling (IJEPC).
Ethics Statement:	This study was conducted in accordance with ethical research standards. All procedures involving human participants were reviewed and approved by the Human Research Ethics Sultan Idris Education University, approval number 2047. Informed consent was obtained from all participants prior to data collection. Participation was voluntary, and respondents were assured of confidentiality and anonymity. The data collected were used solely for academic purposes.
Author Contribution Statement:	All authors contributed significantly to the development of this manuscript. Joelene Janes Lim was responsible for the literature review, drafting, methodology, handled data collection, analysis, and interpretation of results. Nor Hasnida Che Md Ghazali contributed to critical revision of the manuscript and overall supervision of the study. All authors read and approved the final version of the manuscript prior to submission.
Declaration on the Use of Artificial Intelligence (AI)	The author(s) declared that generative AI or AI-assisted technologies were used during the preparation of this manuscript. The following tool(s) were used: Anara, ChatGPT. The tool(s) were used for language editing, grammar improvement and translation. All AI-assisted content was reviewed and verified by the author(s), who take full responsibility for the accuracy, originality, integrity, and final content of the manuscript

International Journal of Education, Psychology and Counseling
IJEPC
EISSN : 0128-164X
Volume 11 Issue 64 (September 2026) PP. 2071-2090

References

Abdul Manap, M. H., Haron, Z., & Othman, N. (2020). Penilaian pelaksanaan program pentaksiran alternatif sekolah rendah (PASR) di sekolah rendah pendidikan khas integrasi. *Jurnal Pendidikan Malaysia*, *45*(Special Issue), 9–16. https://doi.org/10.17576/JPEN-2020-45.01SI-02

Agus, A., Juliadharma, M., & Djamaluddin, M. (2023). Application of the CIPP model in evaluation of the inclusive education curriculum in Madrasah Aliyah. *Nidhomul Haq: Jurnal Manajemen Pendidikan Islam*, *8*(1). https://doi.org/10.31538/ndh.v8i1.2705

Ahmad, N., & Ismail, R. (2020). Challenges in operationalising Alternative Assessment for Primary School as a school-based alternative assessment. *Journal of Education Assessment Research*, *28*(1), 33–47.

Ahmad, N., Ismail, R., & Manap, R. (2021). Validation of teaching, learning, and assessment quality instrument in Malaysian early-childhood context. *International Journal of Education Measurement*, *39*(3), 112–129.

Ang, L. L., & Phoon, H. S. (2025). Behavioural management challenges and support needs in inclusive classrooms in Malaysia. *Asia Pacific Journal of Educators and Education*, *40*(1), 269–292. https://doi.org/10.21315/apjee2025.40.1.14

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Black, P., & Wiliam, D. (2022). Classroom assessment and the improvement of learning. *Assessment in Education: Principles, Policy & Practice*, *27*(1), 1–12. https://doi.org/10.1080/0969594X.2022.2048134

Bujang, M. A., Omar, E. D., Foo, D. H. P., & Hon, Y. K. (2024). Sample size determination for conducting a pilot study to assess reliability of a questionnaire. *Restorative Dentistry & Endodontics*, *49*(1), e3. https://doi.org/10.5395/rde.2024.49.e3

Cheung, G. W. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, *41*(2), 507–548. https://doi.org/10.1007/s10490-023-09871-y

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). Sage Publications.

de Winter, J. C. F., & Dodou, D. (2010). Five-point Likert items: tt test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, *15*(11), 1–12. https://www.researchgate.net/publication/266212127_FivePoint_Likert_Items_t_Test_Versus_Mann-Whitney-Wilcoxon

Dewi, N. K. (2024). A program evaluation study using CIPP model. *International Journal of Learning, Teaching and Educational Research*, *23*(3), 1–15. https://ijlter.org/index.php/ijlter/article/view/12859/0

Examinations Board, Ministry of Education Malaysia. (2016). *Surat Pekeliling Ikhtisas KPM Bil. 3 Tahun 2016: Pelaksanaan Pentaksiran Alternatif Sekolah Rendah (PASR) Murid Berkeperluan Khas mulai tahun 2016*. Lembaga Peperiksaan, Kementerian Pendidikan Malaysia. https://lp.moe.gov.my/index.php/pekeliling-siaran/surat-pekeliling-lembaga-peperiksaan

Examinations Board, Ministry of Education Malaysia. (2026). *Garis Panduan pengurusan Pentaksiran Alternatif Pendidikan Rendah (PAPR) edisi 4 tahun 2026*. Lembaga Peperiksaan, Kementerian Pendidikan Malaysia. https://lp.moe.gov.my/files/papr/2026/Garis%20Panduan%20Pentadbiran%20PAPR%20Edisi%20Keempat.pdf

Ghazali, N. H. C. M., Hamzah, M., Abdullah, N., & Zaini, S. H. (2019). Validation of an instrument to measure the feedback conceptions scale. *International Journal of Academic Research in Business and Social Sciences*, *9*(7). https://doi.org/10.6007/IJARBSS/v9-i7/6091

Gray, D. E. (2021). *Doing research in the real world* (5th ed.). Sage.

Gültürk, O. (2024). Minimum sample size requirements for SEM: A methodological review. *Journal of Quantitative Social Research*, *6*(2), 45–59.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hatem, G., Zeidan, J., Goossens, M., & Moreira, C. (2022). Normality testing methods and the importance of skewness and kurtosis in statistical analysis. *BAU Journal – Science and Technology*, *3*(2), Article 7. https://doi.org/10.54729/KTPE9512

Howard, M. C. (2023). A systematic literature review of exploratory factor analyses in premier journals. *Journal of Business Research*, *163*, 113959. https://doi.org/10.1016/j.jbusres.2023.113959

Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social sciences: The MEASURE approach. *Practical Assessment, Research, and Evaluation*, *26*(1), 1–15. https://doi.org/10.7275/svg4-e671

Kasdiah, M. A. H., Amdan, M. A., & Janius, N. (2024). The impact of school infrastructure on learning among rural areas in Sabah, Malaysia. *UNITAR International University – Education and Humanities Proceedings*, *3*, 224–230.

Khamidah, S., & Sanoto, H. (2025). Evaluating inclusive education implementation using the CIPP model. *Journal of Innovation and Research in Primary Education*, *4*(4). https://doi.org/10.56916/jirpe.v4i4.1874

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.

Knowles, M. S., Holton, E. F., Swanson, R. A., & Robinson, P. A. (2020). *The adult learner* (9th ed.). Routledge.

Koo, M., & Yang, S.-W. (2025). Likert-type scale. *Encyclopedia*, *5*(1), 18. https://doi.org/10.3390/encyclopedia5010018

Levy-Feldman, I. (2025). The role of assessment in improving education and promoting educational equity. *Education Sciences*, *15*(2), 224. https://doi.org/10.3390/educsci15020224

Liu, M., Vu, T., van Atteveldt, N., & Meeter, M. (2023). Testing the reciprocal effect between value of education, time investment, and academic achievement in a large non-Western sample. *Journal of Intelligence*, *11*(7), 133. https://doi.org/10.3390/jintelligence11070133

MCRCR. (2021). *Malaysian Code of Responsible Conduct in Research*. Ministry of Higher Education Malaysia.

Memon, M. A., Ting, H., Cheah, J.-H., Thurasamy, R., Chuah, F., & Cham, T. H. (2020). Sample size for survey research: Review and recommendations. *Journal of Applied Structural Equation Modeling*, *4*(2), 1–19. https://doi.org/10.47263/JASEM.4(2)01

Ministry of Economy Malaysia. (2023). *Malaysia Madani: Ekonomi MADANI–Memperkasa rakyat*. Ministry of Economy Malaysia.

Ministry of Education Malaysia. (2023). *Pelan Pembangunan Pendidikan Malaysia 2013–2025 (PPPM 2013–2025)*. Kementerian Pendidikan Malaysia. https://www.moe.gov.my

Ministry of Education Malaysia. (2026). *Malaysia Education Blueprint 2026–2035*. Ministry of Education Malaysia. https://www.moe.gov.my

Ministry of Finance Malaysia. (2023). *Budget 2023: Belanjawan Madani*. Ministry of Finance Malaysia.

Mulder, M. (2019). Foundations of competence-based vocational education and training. In S. McGrath, M., J. Papier, & R. Suart (Eds.), *Handbook of vocational education and training* (pp. 1–18). Springer. https://doi.org/10.1007/978-3-319-49789-1_65-1

OECD. (2020). *PISA 2018 results: Where all students can succeed*. OECD Publishing. https://doi.org/10.1787/e8d27ed2-en

Ofem, U. J., & Usoro, A. A. (2024). Alternative assessment practices in inclusive classrooms: A review. *International Journal of Educational Research*, *125*, 102341.

Patrick, J., Nordin, M. N., & Amat, S. (2026). The perspective of special education teachers on evidence management in the implementation of the Primary School Alternative Assessment (PAPR). *International Journal of Research and Innovation in Social Science*, *10*(26), 59–68. https://doi.org/10.47772/IJRISS.2026.1026EDU0004

Penang Institute. (2025). *Bridging the digital divide in Malaysian education: Policy briefs on infrastructure, access and equity*. Penang Institute.

Polit, D. F., & Beck, C. T. (2021). *Nursing research: Generating and assessing evidence for nursing practice* (11th ed.). Wolters Kluwer.

Quinney, S. (2021). Inclusive assessment for special educational needs. *Support for Learning*, *36*(4), 562–580. https://doi.org/10.1111/1467-9604.12369

Razali, N. F. Y. M., Mahmud, R., & Aziz, A. (2021). *E-portfolio implementation in Malaysian schools: A need analysis*. International Journal of Academic Research in Progressive Education and Development (Vol.10, Issue 4, pp. 60-70) https://doi.org/10.6007/IJARPED/v10-i4/11842

Sanusi, N., Zulkifli, H., & Hamzah, M. I. (2025). Challenges, opportunities, and effects of alternative assessment approaches in teaching practices: *A systematic review*. International Journal of Educational Research & Evaluation, 14(2), 1105–1113. https://doi.org/10.11591/ijere.v14i2.32283

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students* (8th ed.). Pearson Education Limited.

Stufflebeam, D. L., & Zhang, G. (2022). *The CIPP evaluation model: How to evaluate for improvement and accountability*. Guilford Press.

Sulaiman, L. H., & Tahar, M. M. (2024). *Implementation of universal learning design principles by mainstream teacher in inclusive education programs.* Special Education, 2(1). https://doi.org/10.59055/se.v2i1.15

UNESCO. (2020). *Global education monitoring report 2020: Inclusion and education*. https://unesdoc.unesco.org/ark:/48223/pf0000373718

Ying, X., Freedland, K. E., Powell, L. H., Stuart, E. A., Ehrhardt, S., & Mayo-Wilson, E. (2025). *Determining sample size for pilot trials: A tutorial*. BMJ *2025; 390*, . https://doi.org/10.1136/bmj-2024-083405

Zhao, T., Yu, L., Jiang, X., Yang, S., Zhang, M., Zhang, Y., … Xu, F. (2025). *Construction of a public health practice teaching quality evaluation system based on the CIPP model*. BMC Medical Education, *25*. https://doi.org/10.1186/s12909-025-06943-z