

Development and Validation of the Adolescent Psychological Well-Being Questionnaire (APWQ) for Coastal Fisheries Communities: Content Validity and Rasch Measurement Evidence from Malaysian Secondary School Students

Mohd Radhi Abu Shahim¹, Kamarul Md Shah², Rozita Jayus³, Nurul Nazihah Jamal⁴, Zainudin Bachok⁵

^{1,2,3,4,5} Universiti Malaysia Terengganu, Malaysia.

Abstract

Background. Psychological well-being (PWB) is a central outcome in school-based mental health promotion, yet instruments used in Malaysian intervention research are commonly adopted from Western urban contexts and validated using classical test theory alone, leaving item-level functioning, dimensional structure and sample-instrument targeting unexamined.

Objective. This study developed and validated the Adolescent Psychological Well-Being Questionnaire (APWQ), a 20-item self-report instrument designed for adolescents in coastal fisheries communities, using expert-panel content validation and Rasch measurement theory with secondary school students in Terengganu, Malaysia.

Methods. Content validity was established through a five-member expert panel drawn from four Malaysian universities, rating eight instrument-quality statements on a four-point scale; item-level (I-CVI), scale-level (S-CVI/Ave, S-CVI/UA) and modified kappa indices were computed. Responses from 206 secondary school students were then analysed with the Rasch rating scale model (Winsteps 5.12.0.0), examining category functioning, item and person fit, polarity, unidimensionality, local independence, separation, targeting and differential item functioning (DIF) by gender.

Results. All I-CVI values reached 1.00, with S-CVI/Ave = S-CVI/UA = 1.00 and modified kappa = 1.00 (Excellent) for all eight statements; overall panel endorsement was 99.4% of maximum. The four-category rating scale functioned as intended, with monotonically advancing Andrich thresholds (-1.88, -0.63, +2.51). All 20 items fitted the model (infit MNSQ 0.64–1.27; outfit 0.64–1.34) with positive point-measure correlations (0.38–0.59). Item reliability was 0.97 (separation 5.97) and person reliability 0.85 (separation 2.36; $\alpha = 0.86$). The item difficulty hierarchy was theoretically coherent, running from declarative and normative knowledge (B9, "I understand what bullying is", -1.50 logits) to enacted self-regulatory capability (B6, "I have effective methods for coping with life stress", +1.14 logits). Four items showed gender DIF forming a coherent pattern: male students more readily endorsed self-efficacy and self-knowledge claims (B2, B6, B16), while female students more readily endorsed help-seeking knowledge (B7); only B2 survived Bonferroni correction (contrast 0.83 logits, $p = .0004$). **Conclusion.** The APWQ demonstrates strong content validity and supports interval-level measurement with a theoretically interpretable item hierarchy. Its difficulty range is too narrow and too low for this sample, and item B2 requires revision...

Keywords: psychological well-being; instrument development; Rasch measurement; content validity index; differential item functioning; adolescent mental health; coastal fisheries community; Malaysia.

Introduction

1.1 Background

Adolescence is the developmental window in which most lifetime mental health difficulties first appear, and also the window in which prevention is most tractable. Mental disorders now constitute the leading contributor to disease burden among young people worldwide (Kieling et al., 2024).

The Malaysian evidence is stark. The National Health and Morbidity Survey (NHMS) 2022: Adolescent Health Survey, conducted with 33,523 secondary school students under a nationally representative two-stage stratified cluster design, reported a depressive-symptom prevalence of 26.9% on the PHQ-9 (Institute for Public Health [IPH], 2022; Sahril et al., 2025). The same survey found that fewer than half of adolescents (46.0%) reported the availability of peer support that locating the problem not only in symptom prevalence but in *depleted protective resources*.

This deficit is amplified in geographically peripheral and economically constrained communities. Coastal fisheries communities in Malaysia experience multidimensional deprivation across education, health and financial protection simultaneously (Solaymani & Kari, 2014).

1.3 The Absence of a Context-Specific Instrument

No psychological well-being instrument has been developed and calibrated specifically for adolescents in Malaysian coastal fisheries communities. Researchers and school counsellors working in these settings must therefore choose between generic well-being scales developed for urban or Western populations, whose item content may not reflect the stressors and resources of a fishing household, and unvalidated locally written questionnaires. Neither option supports defensible measurement.

The instrument reported here was constructed to fill that gap. Its item pool was generated from a community needs assessment conducted with children and parents in a coastal fisheries settlement in Terengganu, combined with theory on adolescent well-being, rather than adapted from an existing scale. It is intended for use as a standalone measure in survey research, needs assessment, school counselling screening and the evaluation of well-being interventions in comparable communities.

1.4 The Present Study

The APWQ operationalises adolescent psychological well-being across four content domains, each selected because the literature identifies it as a modifiable determinant of well-being in this age group and because the needs assessment identified it as salient in the study community: mental health and stress awareness, coping strategies and help-seeking, empathy and peer safety, and digital well-being.

This study asks:

Do subject-matter experts endorse the instrument's clarity, language, format and appropriateness for the target group?

Does the four-category rating scale function as intended?

Do items and persons fit the Rasch rating scale model, and is the resulting item hierarchy theoretically interpretable?

Does the instrument satisfy unidimensionality and local independence requirements?

Do separation indices support reliable discrimination?

How well are items targeted to this sample?

Do any items exhibit differential item functioning by gender?

2. Conceptual Framework

2.1 Psychological Well-Being

Psychological well-being has been theorised along two traditions. The hedonic tradition centres on subjective well-being, life satisfaction and affect balance (Diener, 1984). The eudaimonic tradition, operationalised by Ryff (1989), specifies self-acceptance, positive relations, autonomy, environmental mastery, purpose in life and personal growth. Keyes (2002) integrated these into a model of flourishing, demonstrating that the absence of mental illness is an insufficient criterion for mental health the dual-continuum position.

For school-aged populations three qualifications apply. Several eudaimonic dimensions, particularly autonomy and purpose in life, are developmentally emergent, so adult-calibrated instruments frequently mis-target adolescent samples. Well-being at this age is strongly relationally embedded, underscored by the NHMS 2022 peer-support finding (IPH, 2022). And contemporary adolescent well-being is constituted partly in digital space, so a model excluding the online domain is under-specified. This study adopts the Wellness Model (Swarbrick, 2006), conceptualising wellness as a deliberate, self-directed process across interacting dimensions rather than as the absence of pathology.

2.2 The Four Content Domains and Their Evidence Base

The four domains were not chosen for administrative convenience. Each represents a distinct, empirically supported pathway through which adolescent psychological well-being is built or eroded, and each was independently identified as salient in the community needs assessment that generated the item pool.

Mental health and stress awareness addresses recognition. Adolescents cannot seek help for states they cannot name. Mental health literacy interventions produce significant improvements in knowledge and help-seeking and reductions in stigma, with particularly strong effects in Asian settings (Sun et al., 2025); universal educational interventions show convergent literacy gains alongside more variable symptom effects (Yani et al., 2025). School-based universal screening offers near-complete coverage, though primary research remains scarce (Brinley et al., 2024), and recent work specifies mechanisms by which universal programming may produce iatrogenic effects (Foulkes et al., 2026).

Coping strategies and help-seeking addresses the translation of awareness into enacted coping. The transactional model of stress and coping (Lazarus & Folkman, 1984) locates the mechanism at secondary appraisal; social cognitive theory (Bandura, 1997) specifies enactive mastery and vicarious modelling as the acquisition route. Universal social and emotional learning interventions, predominantly skill-rehearsal in design improved skills,

attitudes, behaviours, relationships, school functioning and achievement across 424 studies and 575,361 students in 53 countries (Cipriano et al., 2023).

Empathy and peer safety addresses relational climate. Among Malaysian students, 79.1% report bullying involvement in some role, with bully-victims (48.4%) far outnumbering pure perpetrators (14.4%) or victims (16.3%) (Kutty et al., 2021). Anti-bullying interventions reduce victimisation ($d = -0.14$) and perpetration ($d = -0.07$) (Huitsing et al., 2022), operating primarily through bystander mobilisation, whose proximal target is empathy (Hoffman, 2000).

Digital well-being addresses the environment in which a growing share of adolescent social experience occurs. Cyberbullying perpetration among Malaysian school-going adolescents was estimated at 18.3% in NHMS 2022 (IPH, 2022). Scholarship has moved from protective restriction toward a competence-based conception emphasising student agency, self-regulation and the integration of digital literacy with psychological well-being (Ozdemir & Ulusoy, 2025).

3. Methods

3.1 Design

A two-stage instrument development and validation design was employed: expert-panel content validation followed by cross-sectional field administration analysed with the Rasch rating scale model (Andrich, 1978).

3.2 Instrument

The APWQ comprises 20 self-report items in Malay, administered on a four-point Likert scale (1 = *Sangat Tidak Setuju* / Strongly Disagree; 2 = *Tidak Setuju* / Disagree; 3 = *Setuju* / Agree; 4 = *Sangat Setuju* / Strongly Agree). Items were written by the research team rather than adapted from existing scales. The item pool was generated from two sources: a prior community needs assessment conducted with children and parents in a coastal fisheries settlement in Terengganu, which identified the well-being concerns most salient to that population; and the theoretical and empirical literature. Items map to five constructs nested within the four content domains (Table 1).

Table 1. Item stems, construct mapping and Rasch difficulty calibrations

Item	Malay stem (as administered)	English translation	Construct	Component	Measure (logits)
B1	Saya memahami maksud sebenar kesihatan mental.	I understand the true meaning of mental health.	Mental health awareness	Screening	+0.21
B2	Saya tahu tahap kesihatan mental saya.	I know my own mental health level.	Mental health awareness	Screening	+0.75
B3	Saya menyedari punca tekanan yang berlaku dalam diri saya.	I am aware of the causes of the stress occurring within me.	Stress awareness	Screening	+0.79
B4	Saya menyedari kesan terhadap tekanan kepada kehidupan saya.	I am aware of the effects of stress on my life.	Stress awareness	Screening	+0.54
B5	Saya tahu tanda-tanda tekanan dalam diri saya.	I know the signs of stress within myself.	Stress awareness	Screening	+1.02
B6	Saya ada kaedah yang berkesan dalam menghadapi tekanan hidup.	I have effective methods for coping with life stress.	Coping strategies	Strategies	+1.14
B7	Saya tahu kepada siapa saya boleh mendapatkan bantuan jika mengalami tekanan emosi.	I know whom I can turn to for help if I experience emotional distress.	Help-seeking	Strategies	+0.66
B8	Saya memahami pentingnya menghormati rakan.	I understand the importance of respecting friends.	Empathy	Bullying/Empathy	-1.27

Item	Malay stem (as administered)	English translation	Construct	Component	Measure (logits)
B9	Saya memahami apa itu buli.	I understand what bullying is.	Bullying knowledge	Bullying/Empathy	-1.50
B10	Saya menyedari peranan saya dalam membantu apabila seseorang dibuli.	I am aware of my role in helping when someone is bullied.	Bystander role	Bullying/Empathy	-0.32
B11	Saya tahu tempat yang sesuai untuk mengadu isu buli.	I know the appropriate place to report bullying.	Reporting	Bullying/Empathy	-0.52
B12	Saya tahu cara untuk mengelak dari dibuli.	I know how to avoid being bullied.	Self-protection	Bullying/Empathy	-0.46
B13	Saya tahu cara untuk mengawal diri daripada membuli.	I know how to restrain myself from bullying others.	Self-regulation	Bullying/Empathy	-0.88
B14	Saya faham maksud sebenar kesejahteraan digital.	I understand the true meaning of digital well-being.	Digital well-being	Digital Literacy	+1.03
B15	Saya tahu kaedah menggunakan teknologi dengan selamat dan sihat.	I know how to use technology safely and healthily.	Digital well-being	Digital Literacy	-0.32
B16	Saya mempunyai kaedah untuk mengawal penggunaan digital seperti telefon atau komputer dengan berkesan.	I have effective methods for controlling my digital use, such as phone or computer.	Digital well-being	Digital Literacy	+0.07
B17	Saya sedar bahawa penggunaan teknologi berlebihan boleh memberi kesan kepada mental saya.	I am aware that excessive technology use can affect my mental health.	Digital well-being	Digital Literacy	-0.59
B18	Saya mampu mengawal tempoh penggunaan telefon atau media sosial.	I am able to control the amount of time I spend on my phone or social media.	Digital well-being	Digital Literacy	+0.31
B19	Saya sedar kepentingan keseimbangan antara kehidupan digital dan kehidupan sebenar.	I am aware of the importance of balance between digital life and real life.	Digital well-being	Digital Literacy	-0.23
B20	Saya bersedia mengamalkan penggunaan teknologi digital yang lebih sihat.	I am ready to practise healthier use of digital technology.	Digital well-being	Digital Literacy	-0.44

3.3 Content Validation

Five subject-matter experts attached to four Malaysian universities reviewed the questionnaire before it went into the field. That number meets the minimum of three set by Lynn (1986) and sits inside the three-to-ten range Polit and Beck (2006) treat as workable.

Each reviewer worked through an eight-statement rating form covering format clarity, item meaning, accessibility of the language, legibility, clarity of the instructions, layout, guidance on using the response scale, and suitability for the age group concerned. The form used the same four-point agreement scale as the questionnaire itself.

Three indices were computed from these ratings, following Polit and Beck (2006). I-CVI is the proportion of the five reviewers awarding a 3 or a 4 to a given statement. S-CVI/Ave is the mean of the eight I-CVI values, and S-CVI/UA the proportion of statements on which all five reviewers awarded a 3 or a 4. Because agreement of this kind is partly a matter of chance, a modified kappa was also calculated: $\kappa^* = (I-CVI - P_c) / (1 - P_c)$, with $P_c = [N! / (A!(N - A)!)] \times 0.5^N$. Following Polit et al. (2007), κ^* above 0.74 was read as excellent, 0.60–0.74 as good, and 0.40–0.59 as fair. One consequence of the panel size is worth stating plainly: at five experts, Lynn's (1986) criterion admits nothing short of unanimous endorsement, so a single reservation puts a statement below the acceptable I-CVI.

3.4 Participants and Field Administration

Questionnaires were completed by 206 secondary school students from a coastal fisheries community in Terengganu, on the east coast of Peninsular Malaysia. Of these, 89 were boys (43.2%) and 117 girls (56.8%).

3.5 Ethical Considerations

Written informed consent was obtained from parents or legal guardians of all adolescent participants, and written assent was obtained from the adolescents themselves. Adult participants in the parent focus group and the needs survey provided written informed consent. Participants were informed of the voluntary nature of participation, of their right to withdraw at any point without consequence, and of the procedures used to protect confidentiality. All records were de-identified prior to analysis, and audio recordings were stored separately from transcripts on password-protected media.

3.6 Rasch Analysis

Analyses were conducted in Winsteps 5.12.0.0 using joint maximum likelihood estimation under the rating scale model, with a single set of Andrich thresholds estimated across items given the common response format. Criteria are set out in Table 2.

Table 2. Rasch analysis criteria and sources

Property	Index	Criterion	Source
Category functioning	Observed count per category	≥ 10 observations	Linacre (2002)
	Average measures	Advance monotonically	Linacre (2002)
	Category outfit MNSQ	< 2.00	Linacre (2002)
	Andrich thresholds	Advance monotonically; 1.4–5.0 logits apart	Linacre (2002)
Item polarity	Point-measure correlation	Positive; ≥ 0.30 preferred	Boone et al. (2014)
Item/person fit	Infit & outfit MNSQ	0.60–1.40 (rating scales)	Bond & Fox (2015)
	ZSTD	–2.0 to +2.0 (secondary; sample-sensitive)	Linacre (2002)
Unidimensionality	Raw variance explained	Compare observed vs expected	Linacre (2012)
	Eigenvalue, 1st contrast	< 2.0 preferred	Linacre (2012)
	Disattenuated cluster correlation	≈ 1.0 indicates common dimension	Linacre (2012)
Local independence	Standardised residual correlation	< 0.30	Linacre (2012)
Reliability	Item / person reliability	> 0.80 good; > 0.90 excellent	Fisher (2007)
	Separation	> 2.0 acceptable	Fisher (2007)
	Strata = $(4G + 1)/3$	≥ 3 distinguishable groups	Wright & Masters (1982)
Targeting	Person mean – item mean	Within ± 0.50 logits	Boone et al. (2014)

Property	Index	Criterion	Source
DIF	Rasch-Welch t	$p < .05$; Bonferroni $\alpha = .05/20 = .0025$	—
	DIF contrast (ETS class)	<0.43 negligible; $0.43-0.64$ moderate; ≥ 0.64 large	Zwick et al. (1999)

4. Results

4.1 Content Validity

Expert ratings are reported in Table 3. Of 40 ratings, 39 were at the scale maximum; a single expert (EV5) rated statement 2 at 3 (*Setuju*).

Table 3. Expert panel ratings and content validity indices (5 experts, 8 statements)

#	Statement (English)	EV1	EV2	EV3	EV4	EV5	Mean	Experts rating 3–4	I-CVI	κ^*	Interpretation
1	The format of this questionnaire is clear	4	4	4	4	4	4.00	5	1.00	1.00	Excellent
2	The meaning of each item is clear	4	4	4	4	3	3.80	5	1.00	1.00	Excellent
3	The language used is easy to understand	4	4	4	4	4	4.00	5	1.00	1.00	Excellent
4	The font size is appropriate and easy to read	4	4	4	4	4	4.00	5	1.00	1.00	Excellent
5	The instructions given are clear	4	4	4	4	4	4.00	5	1.00	1.00	Excellent
6	The line spacing is appropriate	4	4	4	4	4	4.00	5	1.00	1.00	Excellent
7	The guide for the measurement scale is clear	4	4	4	4	4	4.00	5	1.00	1.00	Excellent
8	This questionnaire is appropriate for the target study group	4	4	4	4	4	4.00	5	1.00	1.00	Excellent
	Overall						3.98		S-CVI/Ave = 1.00		

Note. S-CVI/UA = 1.00 (8/8 statements achieved universal agreement). P_c for $N = 5$, $A = 5 = 0.031$. Total panel endorsement = $159/160 = 99.4\%$ of maximum possible.

All I-CVI values reached 1.00, satisfying Lynn's (1986) requirement of 100% agreement for a five-member panel. S-CVI/Ave and S-CVI/UA both reached 1.00, exceeding the 0.90 criterion. Modified kappa was 1.00 for every statement, classified Excellent (Polit et al., 2007). No statement required revision on content-validity grounds, and the panel confirmed the instrument's appropriateness for the target group (statement 8, unanimous).

4.2 Rating Scale Category Functioning

The four-category scale functioned as intended (Table 4). All categories exceeded the minimum observation threshold, although Category 1 was sparsely used (122 observations, 3.0%). Observed average measures advanced across the four categories, from -0.25 at the lowest to $+2.34$ at the highest. Category outfit MNSQ values stayed below 2.00 throughout, and the Andrich thresholds advanced in order (-1.88 , -0.63 , $+2.51$). The categories were therefore used in the intended sequence, and none introduced unmodelled noise.

Table 4. Summary of category function (N = 206; 4,120 observations)

Category	Count	%	Observed average	Expected	Infit MNSQ	Outfit MNSQ	Andrich threshold	Category measure
1 Sangat Tidak Setuju	122	3	-0.25	-0.58	1.22	1.32	—	(-3.15)
2 Tidak Setuju	645	16	0.04	0.17	0.89	0.88	-1.88	-1.28
3 Setuju	2,282	55	1.15	1.13	0.94	0.94	-0.63	0.99
4 Sangat Setuju	1,071	26	2.34	2.34	1.02	1.01	+2.51	(3.64)

Two qualifications apply. First, the separation between the first and second thresholds was 1.25 logits. Although this is slightly below the recommended 1.4-logit distance, it suggests only a modest overlap between Categories 2 and 3 rather than a serious problem with the response scale. In practical terms, respondents may not distinguish these two categories as clearly as intended. Second, three items B6, B10, and B17 showed a minor reversal in the average ability estimates: respondents selecting Category 2 did not demonstrate a higher average ability than those selecting Category 1. However, Category 1 was selected by very few respondents for these items (B6: n=13; B10: n=4; B17: n=2). Given these small frequencies, the reversals are more plausibly attributable to unstable estimates from sparse observations than to substantive category disordering.

4.3 Item Polarity and Fit

The point-measure correlations for all 20 items were positive, ranging from 0.38 for B8 to 0.59 for B3. As every value exceeded the minimum criterion of 0.30, the items were considered to align positively with the intended construct. This indicates that higher responses on each item were generally associated with higher levels of the measured attribute.

Item-fit results also supported retention of all items. Both infit and outfit mean-square statistics were within the acceptable range of 0.60 to 1.40 for rating-scale data, as presented in Table 5. Infit MNSQ values ranged from 0.64 for B19 to 1.27 for B7 and B11. Outfit MNSQ values ranged from 0.64 for B19 to 1.34 for B7. These findings indicate that none of the items displayed excessive misfit or response unpredictability; therefore, all 20 items were retained for subsequent analysis.

Table 5. Item measures and fit statistics, ordered by difficulty

Rank	Item	Abbreviated stem	Measure	SE	Infit MNSQ	Outfit MNSQ	PTMEA
1	B6	Have effective coping methods	1.14	.11	0.88	0.91	.58
2	B14	Understand meaning of digital well-being	1.03	.11	0.91	0.93	.49
3	B5	Know signs of stress in myself	1.02	.11	1.01	1.03	.53
4	B3	Aware of causes of my stress	0.79	.11	0.92	0.95	.59
5	B2	Know my own mental health level	0.75	.11	1.26	1.29	.46
6	B7	Know whom to turn to for help	0.66	.11	1.27	1.34	.49
7	B4	Aware of effects of stress on my life	0.54	.11	1.19	1.16	.50
8	B18	Can control phone/social media time	0.31	.12	1.08	1.12	.47
9	B1	Understand meaning of mental health	0.21	.12	0.93	0.90	.47
10	B16	Have methods to control digital use	0.07	.12	0.93	0.91	.54
11	B19	Aware of digital/real-life balance	-0.23	.12	0.64	0.64	.57

Rank	Item	Abbreviated stem	Measure	SE	Infit MNSQ	Outfit MNSQ	PTMEA
12	B10	Aware of my role in helping bullied peers	-0.32	.12	0.96	0.96	.47
13	B15	Know safe and healthy technology use	-0.32	.12	0.83	0.78	.58
14	B20	Ready to practise healthier digital use	-0.44	.13	0.87	0.87	.55
15	B12	Know how to avoid being bullied	-0.46	.13	0.95	0.91	.55
16	B11	Know where to report bullying	-0.52	.13	1.27	1.23	.50
17	B17	Aware excessive tech affects mental health	-0.59	.13	1.01	0.98	.48
18	B13	Know how to restrain myself from bullying	-0.88	.13	0.99	0.97	.55
19	B8	Understand importance of respecting friends	-1.27	.14	1.17	1.13	.38
20	B9	Understand what bullying is	-1.50	.14	0.94	0.92	.47
	Mean		0.00	.12	1.00	1.00	
	SD		0.76	.01	0.16	0.17	

Although five items recorded ZSTD values outside the conventional range of ± 2.0 , their MNSQ values remained within the acceptable limits. This pattern is not unusual, particularly when the sample size is relatively large. ZSTD is highly sensitive to sample size because it evaluates departures from an ideal model-fit expectation. Consequently, even small and practically unimportant deviations may become statistically significant. MNSQ was therefore given greater interpretive priority because it reflects the practical magnitude of misfit and is less affected by sample size (Linacre, 2002).

Two items nevertheless warrant further consideration. Item B7 produced the highest outfit statistic (1.34), indicating relatively greater response unpredictability than the other items. This suggests that some respondents reported knowing whom to approach for assistance in a manner that was not fully consistent with their overall level of well-being. However, because its outfit value remained within the accepted range, the item was retained. In contrast, B19 produced low infit and outfit MNSQ values of 0.64, indicating overfit. Responses to this item were more predictable than expected under the Rasch model and followed a pattern resembling a Guttman scale. Such overfit may indicate that B19 overlaps conceptually with nearby items assessing digital awareness. Nevertheless, the statistic remained within the acceptable range, so the item was retained while its potential redundancy should be considered during future instrument refinement.

4.4 The Item Difficulty Hierarchy

The item hierarchy extended across 2.64 logits and showed a pattern consistent with the theoretical expectations underlying the construct. This spread indicates that the instrument captured varying levels of item difficulty, from relatively easy to more demanding indicators. The coherence of the hierarchy with the proposed construct provides the principal evidence of construct validity in this study.

The easiest items to endorse are declarative and normative. B9 ("I understand what bullying is", -1.50) and B8 ("I understand the importance of respecting friends", -1.27) require only general conceptual knowledge and assent to a social norm. B13 ("I know how to restrain myself from bullying", -0.88) and B17 ("I am aware that excessive technology use can affect my mental health", -0.59) similarly index awareness of widely taught propositions.

The hardest items require enacted self-regulatory capability and introspective self-knowledge. B6 ("I have effective methods for coping with life stress", +1.14) asserts a functioning coping repertoire. B5 ("I know the signs of stress within myself", +1.02), B3 (+0.79) and B2 ("I know my own mental health level", +0.75) all require the respondent to turn attention inward and claim accurate self-perception.

The continuum therefore runs from general declarative knowledge → normative endorsement → self-directed awareness → enacted capability. This is precisely the progression predicted by mental health literacy theory, in which recognition precedes management (Jorm et al., 1997), and by social cognitive theory, in which knowledge and self-efficacy for enactment are distinct and sequentially acquired (Bandura, 1997). That an instrument

developed inductively from community needs data should reproduce this theoretical ordering empirically is strong evidence that it measures what it claims to. Construct-level means (Table 6) make the pattern explicit.

Table 6. Mean item difficulty by construct

Construct	Items	n	Mean measure	Range
Coping strategies & help-seeking	B6–B7	2	+0.90	+0.66 to +1.14
Stress awareness	B3–B5	3	+0.78	+0.54 to +1.02
Mental health awareness	B1–B2	2	+0.48	+0.21 to +0.75
Digital well-being	B14–B20	7	−0.02	−0.59 to +1.03
Empathy & anti-bullying	B8–B13	6	−0.83	−1.50 to −0.32

The empathy and peer-safety construct sits 0.83 logits below the instrument mean and 1.73 logits below the coping construct. Students in this sample endorse bullying-related knowledge readily and near-universally, while reporting substantially less confidence in stress self-awareness and coping capability.

Read as a substantive finding rather than a psychometric statistic, this is informative in its own right: in this population, anti-bullying awareness appears already well established, plausibly reflecting sustained national and school-level campaigns whereas the capacity to recognise and manage one's own stress is where adolescents report the greatest deficit. For practitioners using the APWQ as a needs-assessment tool, that is the substantive signal the instrument delivers, and it has implications for where limited counselling resources are directed.

One item merits separate note. B14 ("I understand the true meaning of digital well-being", +1.03) is the second-hardest item overall, while the other six digital items average −0.20. The likely explanation is not that digital well-being is a hard state to achieve but that the digital wellbeing is an unfamiliar formal term. B14 may therefore be measuring vocabulary comprehension rather than the construct, a source of construct-irrelevant difficulty warranting review.

4.5 Person Fit

Across the 204 non-extreme respondents, the mean infit and outfit MNSQ values were both 1.00, with a standard deviation of 0.60. Overall, this indicates that the response data showed an acceptable level of consistency with Rasch model expectations. Seventeen respondents (8.3%) recorded outfit MNSQ values of at least 2.00 together with ZSTD values of at least 2.0. Although this proportion was slightly higher than the approximately 5% that may be expected by chance, inspection of the response patterns suggested that the misfit was mainly caused by isolated unexpected responses rather than systematic inconsistency. Person-fit statistics are intended to identify response patterns that are improbable under the Rasch model, including random or unusually extreme responses. For example, respondent 178 selected a low response category for B17 despite an expected score of 3.29, producing a standardised residual of −4.13. Similarly, respondent 70 responded unexpectedly to B11, with an observed score below the expected value of 3.27 and a residual of −4.09. These isolated deviations are consistent with occasional unexpected endorsements rather than evidence that the respondents answered randomly throughout the instrument.

A further 20 respondents showed outfit values below 0.40, indicating response patterns that were more predictable than the model would anticipate. Such low outfit values may reflect highly uniform response behaviour, although this pattern alone does not establish whether respondents were carefully consistent or simply using a narrow response strategy. MNSQ values are generally interpreted as the more practically meaningful fit indicator, whereas ZSTD can be more sensitive to the number of observations. Because the misfitting respondents represented a relatively small proportion of the sample, appeared across different levels of the latent measure, and were not linked to a single problematic item, all 206 respondents were retained for the subsequent analyses.

4.6 Unidimensionality and Local Independence

Table 7. Principal components analysis of standardised residuals

Component	Eigenvalue	Observed %	Expected %
Total raw variance in observations	30.92	100.0	100.0
Raw variance explained by measures	10.92	35.3	35.6
Raw unexplained variance	20.00	64.7	64.4
1st contrast	2.41	7.8	—

Raw variance explained by measures was 35.3% against a model-expected 35.6%, a near-exact correspondence, which is more informative than the absolute percentage since expected variance is itself a function of measure dispersion. Essential unidimensionality (Rasch/common variance) was 54.4%.

4.6.1 The First Contrast Is Substantive, Not Artefactual

The first contrast eigenvalue of 2.41 exceeds the conventional 2.0 threshold. With the item content now available, its composition is interpretable and the interpretation is not the one that item labels alone would suggest. Local independence was supported. The largest standardised residual correlations were 0.30 (B3–B4 and B9–B10), 0.29 (B4–B5) and 0.28 (B3–B5), all at or below the 0.30 criterion and far below the 0.70 level indicating serious dependency. Notably, the largest residual correlations occur within the same construct pairs identified in the contrast analysis, consistent with the multi-component reading above.

4.7 Separation and Reliability

Table 8 summarises the Rasch measurement results for the 206 respondents and 20 items. The mean person measure was 1.29 logits, compared with the fixed item mean of 0.00 logits. This difference indicates that, on average, respondents were located above the difficulty level of the item set. Person measures showed moderate variation, with a standard deviation of 1.19 logits and values ranging from –1.08 to 6.94 logits. The average standard error of person estimates was 0.40 logits.

The items showed a smaller spread than the respondents, with a standard deviation of 0.76 logits. Item difficulty estimates ranged from –1.50 logits to 1.14 logits, confirming that the item pool covered a narrower segment of the construct than the respondent distribution. The mean item standard error was low at 0.12 logits, indicating stable item-location estimates.

Cronbach's α for raw scores was 0.86. Person real separation was 2.36, with real reliability of .85. These values suggest that the instrument can distinguish respondents into approximately 3.48 statistically distinct levels of the measured construct. The item results showed stronger separation and reliability, with a real separation index of 5.97 and real reliability of .97. This indicates that the sample was sufficiently large and diverse to confirm the relative ordering of the 20 items, yielding approximately 8.29 distinct item-difficulty strata.

The real RMSE was .47 logits for persons and .12 logits for items. Overall, the findings demonstrate satisfactory person reliability and very strong item reliability, while also showing that the item set was less extensive than the range of respondent ability measures.

Table 8. Summary statistics for persons and items

Statistic	Persons (all, N = 206)	Items (n = 20)
Mean measure	1.29	0.00
SD	1.19	0.76
Mean SE	.40	.12
Maximum	6.94	1.14
Minimum	–1.08	–1.50
Real RMSE	.47	.12
Real separation	2.36	5.97
Real reliability	.85	.97
Strata	3.48	8.29

4.8 Differential Item Functioning by Gender

Differential item functioning (DIF) was examined to determine whether individual items operated differently for male and female students after accounting for their overall level on the measured construct. The analysis compared 89 male students with 117 female students. Four items showed statistically significant DIF at the uncorrected significance level, as reported in Table 9.

Table 9. Items showing statistically significant DIF by gender

Item	Abbreviated stem	Male DIF measure	Female DIF measure	Contrast	Joint SE	t	df	p	Mantel χ^2	p	ETS
B2	Know my own mental health level	0.26	1.09	–0.83	.23	–3.61	179	.0004	6.08	.0137	C (large)

Item	Abbreviated stem	Male DIF measure	Female DIF measure	Contrast	Joint SE	t	df	p	Mantel χ^2	p	ETS
B16	Have methods to control digital use	-0.28	0.32	-0.60	.24	-2.45	182	.0151	7.12	.0076	B (moderate)
B6	Have effective coping methods	0.81	1.38	-0.56	.22	-2.55	183	.0117	6.41	.0113	B (moderate)
B7	Know whom to turn to for help	0.96	0.43	+0.52	.23	+2.30	192	.0228	5.14	.0233	B (moderate)

Note. Negative contrast = item easier for male students. The remaining 16 items showed non-significant DIF with contrasts from 0.00 to 0.49. Bonferroni-corrected $\alpha = .05/20 = .0025$.

Applying Bonferroni correction, only B2 survives ($p = .0004$). Its contrast of 0.83 logits also places it in ETS Category C, and it is the only item meeting both magnitude and corrected-significance criteria. B2 additionally showed elevated misfit (infit 1.26, outfit 1.29).

The four items form a coherent and interpretable pattern. Three of them, B2 ("I know my own mental health level"), B6 ("I have effective methods for coping with life stress") and B16 ("I have effective methods for controlling my digital use") are all self-efficacy or self-knowledge claims, and all three were easier for male students to endorse. The fourth, B7 ("I know whom I can turn to for help"), is a help-seeking item, and it was easier for female students.

This maps onto a well-documented gendered pattern in adolescent self-report: boys report higher self-assessed competence and lower help-seeking orientation, while girls report greater willingness to seek help alongside more conservative self-appraisal. The finding is notable because female adolescents typically score *higher* on measured mental health literacy; that they were nonetheless *less* likely to endorse B2 at equivalent overall well-being suggests B2 taps self-assessed certainty about one's own mental state rather than literacy as such a construct validity concern independent of the DIF.

B6, B16 and B7 do not survive correction, and with four of twenty items reaching uncorrected significance where one is expected by chance, individual results warrant caution. The *pattern* across them is nonetheless consistent enough to justify examining self-efficacy-worded items as a class in the next administration, rather than treating each as an isolated flag.

5. Discussion

5.1 Convergent and Divergent Evidence from Two Validation Stages

The content validation returned a perfect result: every I-CVI, S-CVI/Ave, S-CVI/UA and modified kappa at 1.00, with 99.4% overall panel endorsement. Expert panels evaluate whether an instrument *is well made*, clear, legible, appropriately worded, suitable for its audience. Rasch analysis evaluates how an instrument *behaves when administered* whether its items are positioned where respondents actually sit, whether they function equivalently across groups, whether the response structure supports the inferences drawn from it. It is worth noting that the panel's single sub-maximal rating fell on statement 2, "the meaning of each item is clear", the one statement addressing item-level content rather than presentation.

5.2 Construct Validity from the Item Hierarchy

The strongest validity evidence in this study is the difficulty hierarchy. An instrument built inductively from community needs data, without reference to an existing scale, produced an ordering that runs from declarative knowledge through normative endorsement to introspective awareness and enacted capability, the progression that mental health literacy theory and social cognitive theory both predict.

This matters more than the reliability coefficients. High reliability shows that items hang together; a theoretically ordered hierarchy shows that they hang together along the continuum the construct specifies.

5.3 Multidimensionality by Design

The first residual contrast distinguished items assessing introspective awareness of stress from items assessing knowledge of anti-bullying actions. These item sets correspond to two of the instrument's four intended content domains. This pattern is therefore more appropriately interpreted as a consequence of the instrument's deliberate

breadth than as evidence of a measurement defect. Residual contrasts can reflect substantive content strands within a broader construct, and their practical importance should be judged by whether they warrant separate measures rather than by the contrast alone.

A total score can be retained as an overall indicator of student well-being. This interpretation is supported by the acceptable item-fit results, consistently positive point-measure correlations, and close agreement between observed and model-expected variance. Together, these results indicate that the items function sufficiently well as indicators of a common general construct.

However, domain-level scores should be prioritised where the instrument is used for screening, needs assessment, or formative intervention planning. The 1.73-logit difference between the peer-safety and coping domain means indicates a meaningful difference in students' relative standing across these areas. A single composite score would conceal that distinction: two students with identical overall scores could have notably different profiles of peer safety and coping.

5.4 The Targeting Problem and Its Consequences

The instrument is mistargeted by 1.29 logits, leaving the upper 3.8 logits of the person distribution without item coverage. Three consequences follow.

Measurement precision is unequal across the sample. Respondents in the upper half are measured with substantially larger standard errors. Any subgroup analysis involving high-well-being students rests on noisier estimates.

The construct-level pattern indicates where new items are most needed. The anti-bullying construct sits at -0.83 logits, meaning students in this population already endorse bullying knowledge close to universally. Adding difficulty in that construct is unlikely to be productive. The coping and stress-awareness constructs, at $+0.90$ and $+0.78$, are closer to where students actually sit and are the natural place to extend upward and are also the domains in which adolescents in this population report the greatest room for growth.

A ceiling of this kind is a predictable consequence of the development pathway. Items generated to be comprehensible and acceptable to students in a specific community, the correct priority at design stage will tend to cluster in the readily endorsed range. The remedy is not to abandon that pathway but to complete it with a second generation of items pitched deliberately higher.

5.5 Differential Item Functioning

Following adjustment for multiple DIF comparisons, 19 of the 20 items showed no statistically significant gender-related DIF. This provides evidence that the instrument functions comparably for male and female students and supports pooling both groups in mixed-gender analyses. DIF decisions should, however, consider statistical significance together with the magnitude of the DIF contrast and its substantive relevance (Linacre, 2012; Hagquist & Andrich, 2017).

Item B2 was the sole exception. Its DIF contrast of 0.83 logits is substantial relative to the instrument's total item span of 2.64 logits, representing approximately one-third of the available measurement range. When considered alongside its elevated fit statistics and potential interpretive concerns, this result suggests that B2 should be revised rather than retained in its current form. An item with a large DIF contrast may operate differently for male and female students even when both groups have comparable overall levels of well-being (Linacre, 2012).

The overall pattern of DIF is also substantively meaningful. Male students appeared more likely to endorse statements reflecting self-efficacy, whereas female students more readily endorsed items related to knowledge of help-seeking. Previous research has reported that gender differences in adolescent self-efficacy and help-seeking may occur, although the direction and magnitude of these differences vary across settings and domains (Spence et al., 2010; Sears & MacIntyre, 2023; Martínez et al., 2021).

For future instrument refinement, statements framed as self-evaluations of competence, such as "I have effective methods for ..." or "I know my own ...," should be reviewed carefully because they may be influenced by gendered norms of self-presentation. Items expressed as observable behaviours may be less vulnerable to this source of variation. For example, the statement "I have effective methods for managing stress" could be revised as: "When I feel stressed, I use a strategy that helps me calm down or deal with the situation." The revised item should subsequently be evaluated using item-fit and DIF analyses in an independent sample.

5.6 Implications for Instrument Revision

Generate 5–8 new items targeting $+1.5$ to $+4.0$ logits, concentrated in the coping and stress-awareness domains. This is the priority action, addressing targeting, extending measurement range and testing the contrast interpretation simultaneously.

Revise or remove B2. Consider replacing self-assessed certainty ("I know my own mental health level") with a behavioural indicator less exposed to gendered self-presentation.

Review B14 for vocabulary artefact. If digital wellbeing is unfamiliar, either define it in the item stem or replace it with concrete description.

Review B19 for redundancy (overfit 0.64).

Monitor B7 (outfit 1.34) and re-examine its wording for ambiguity about what "knowing whom to turn to" requires.

Test behaviourally worded alternatives for the self-efficacy items (B2, B6, B16) and re-run DIF.

Re-calibrate on an independent sample before use as a primary outcome measure and note that if the five-point scale revision is adopted, all threshold calibrations reported here require re-estimation.

5.7 Limitations

The sample is drawn from a single fisheries community, so calibrations are specific to that population. The cross-sectional design permits no assessment of test–retest stability or of measurement invariance over time, the latter directly relevant given intended pre–post use. DIF was examined for gender only; age, school and form-level DIF remain unexamined, and with 206 respondents, individual DIF estimates carry appreciable standard error (joint SE 0.22–0.28). The person misfit rate of 8.3%, together with 20 respondents showing marked overfit, raises the possibility of response-style effects not modelled here. The content validation covers instrument quality rather than item-level relevance, and its ceiling limits its discriminating value. No external criterion measure was administered, so convergent and discriminant validity evidence is absent. Finally, the unidimensional rating scale model was applied to an instrument that the dimensionality analysis suggests is multi-component; results should be read with that specification choice in view.

6. Conclusion

This study developed and validated the Adolescent Psychological Well-Being Questionnaire (APWQ), a 20-item self-report instrument for adolescents in Malaysian coastal fisheries communities — a population for which no context-specific well-being measure previously existed. A five-member expert panel returned unanimous endorsement across all indices ($I-CVI = S-CVI/Ave = S-CVI/UA = \kappa^* = 1.00$). Rasch analysis confirmed that all 20 items fit the model, the four-category scale functions as intended, items are locally independent, and item calibrations are highly replicable (item reliability 0.97). Most importantly, the item difficulty hierarchy reproduced the theoretically predicted progression from declarative knowledge to enacted self-regulatory capability, supplying construct validity evidence that reliability indices cannot.

Three findings qualify this. The instrument is mistargeted for this sample by 1.29 logits, with approximately 3.8 logits of the upper person continuum uncovered, a limitation with direct consequences for detecting change over time. The first residual contrast reflects the instrument's deliberate multi-domain design, separating introspective stress awareness from anti-bullying knowledge, so composite scores should be reported as such rather than as a unidimensional trait. And item B2 exhibits large gender DIF within a broader pattern in which male students more readily endorse self-efficacy claims and female students more readily endorse help-seeking knowledge.

The APWQ is fit for cross-sectional research, needs assessment and counselling screening with these caveats, while its use as a pre–post outcome measure should await the addition of higher-difficulty items. The revision path is clear and bounded. Reporting the targeting and dimensionality limitations transparently, rather than presenting perfect content validity and strong reliability alone, is what allows the next generation of the instrument to be built on evidence rather than assumption..

References

1. Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573.
2. Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.
3. Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.
4. Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.
5. Brinley, S. K., Tully, L. A., Carl, T., McLean, R. K., Cowan, C. S. M., Hawes, D. J., Dadds, M. R., & Northam, J. C. (2024). Universal child mental health screening for parents: A systematic review of the evidence. *Prevention Science*, 25(5), 798–812. <https://doi.org/10.1007/s11121-024-01693-8>
6. Cipriano, C., Strambler, M. J., Naples, L. H., Ha, C., Kirk, M., Wood, M., Sehgal, K., Zieher, A. K., Eveleigh, A., McCarthy, M., Funaro, M., Ponnock, A., Chow, J. C., & Durlak, J. (2023). The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions. *Child Development*, 94(5), 1181–1204. <https://doi.org/10.1111/cdev.13968>
7. Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, 95(3), 542–575.
8. Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, 21(1), 1095.
9. Foulkes, L., Guzman Holst, C., & Andrews, J. L. (2026). Potential harm from universal school-based mental health interventions: Candidate mechanisms and future directions. *Current Opinion in Psychology*, 67, 102196. <https://doi.org/10.1016/j.copsyc.2025.102196>

10. Hoffman, M. L. (2000). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.
11. Huitsing, G., et al. (2022). What works for whom in school-based anti-bullying interventions? An individual participant data meta-analysis. *Prevention Science*. <https://doi.org/10.1007/s11121-022-01387-z>
12. Institute for Public Health. (2022). Technical report: National Health and Morbidity Survey (NHMS) 2022 — Adolescent Health Survey, Malaysia. National Institutes of Health, Ministry of Health Malaysia. https://iku.gov.my/images/nhms-2022/Report_Malaysia_nhms_ahs_2022.pdf
13. Jorm, A. F., Korten, A. E., Jacomb, P. A., Christensen, H., Rodgers, B., & Pollitt, P. (1997). "Mental health literacy": A survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. *Medical Journal of Australia*, 166(4), 182–186.
14. Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, 43(2), 207–222.
15. Kieling, C., et al. (2024). Worldwide prevalence and disability from mental disorders across childhood and adolescence. [confirm] — cited via Balasooriya Lekamge et al. (2025), *Journal of Youth and Adolescence*, 54(2), 271–289.
16. Kutty, F. M., et al. (2021). Bullying and its associated individual, peer, family and school factors: Evidence from Malaysian national secondary school students. [confirm] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8297093/>
17. Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer.
18. Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85–106.
19. Linacre, J. M. (2012). *Winsteps® Rasch measurement computer program user's guide*. Winsteps.com.
20. Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–385.
21. Ozdemir, S., & Ulusoy, M. (2025). The relationship between school-age students' literacy skills and digital well-being: A systematic review. *BMC Psychology*. <https://doi.org/10.1186/s40359-025-03573-4>
22. Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497.
23. Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459–467.
24. Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, 57(6), 1069–1081.
25. Sahril, N., et al. (2025). Risk and protective factors associated with depressive symptoms among school-going adolescents in Malaysia: A cross-sectional study. *BMC Psychiatry*. <https://doi.org/10.1186/s12888-025-07319-9>
26. Solaymani, S., & Kari, F. (2014). Poverty evaluation in the Malaysian fishery community. *Ocean & Coastal Management*, 95, 165–175.
27. Sun, G., Wang, C., & Zhang, J. (2025). Effectiveness of mental health literacy interventions for adolescents: A systematic review and meta-analysis. *SAGE Open*. <https://doi.org/10.1177/21582440251327445>
28. Swarbrick, M. (2006). A wellness approach. *Psychiatric Rehabilitation Journal*, 29(4), 311–314.
29. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.
30. Yani, D. I., et al. (2025). The effects of universal educational interventions in improving mental health literacy, depression, and anxiety among adolescents: A systematic review and meta-analysis. *International Journal of Mental Health Nursing*. <https://doi.org/10.1111/inm.13494>
31. Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36(1), 1–28.

.

.

..