KONSELING: Jurnal Ilmiah Penelitian dan Penerapannya
Vol.8, No.1, Oktober 2026, pp. 38-47
e-ISSN: 2686-2875
https://journal.ilininstitute.com/konseling
DOI: 10.31960/konseling.v8i1.3538
Received on 22/08/2026; Revised on 21/09/2026; Accepted on 19/10/2026; Published on: 29/10/2026



Development of an Academic Stress Scale for Hight School Students: A Rasch Analysis

Fitri Yulianti*)[1], Elwas Berdha Krismona[2], M. Fiqri Syahril[3]
[1]Guidance and Counseling, Universitas Lampung, Indonesia
[2]Guidance and Counseling, Universitas Sebelas Maret, Indonesia
[3]Gudance and Counseling, Universitas Negeri Makassar, Indonesia
*)Corresponding author, E-mail: fitriyulianti.fy@fkip.unila.ac.id

Abstract. Academic stress is an important concern in education, yet the measurement properties of academic stress scales require careful examination to establish whether their items and response categories function adequately and whether item responses differ across student groups. This study aimed to develop an academic stress scale and evaluate its measurement properties using the Rasch model. Data were collected from 322 senior and vocational high school students in Indonesia (190 females and 132 males). The initial 22-item instrument assessed emotional, cognitive, and behavioral reactions to academic demands. Four items did not meet the specified item-fit criteria, resulting in an 18-item scale. The revised four-category response scale showed increasing observed average measures, although one threshold difference remained below the recommended range. The scale demonstrated good internal consistency (Cronbach's alpha = 0.84) and high item reliability (0.98), and the dimensionality criteria were met. Gender-related differential item functioning (DIF) was identified in eight of the 18 retained items, indicating potential differences in item functioning between female and male students. These findings highlight the need to examine DIF more closely before interpreting gender-group comparisons. Further evaluation using more diverse student samples and additional evidence is needed to establish the scope of the scale's applicability.

Keyword: academic stress, instrument development, validation, rasch model.

This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2026 by author.

Introduction

Academic stress is an important issue in education because academic demands can become a source of pressure when students face examinations, homework, grade targets, and expectations regarding academic achievement (Wuthrich et al., 2020) . Such pressure is not limited to assessment activities but may also arise from increasing workloads, expectations of independent learning, and demands from students' families, schools, and peers (Wuthrich et al., 2020) . Differences in students' levels of stress may also be associated with individual characteristics, such as a tendency toward anxiety, self-efficacy, coping strategies, and resilience (Wuthrich et al., 2020). Other factors associated with academic stress include optimism, achievement motivation, and procrastination, while social support is a relevant external factor (Yusuf & Yusuf, 2020) . Thus, academic stress should be understood as an experience shaped by educational demands, individual characteristics, and students' social environments (Wuthrich et al., 2020; Yusuf & Yusuf, 2020) . The significance of this issue is also reflected in studies linking academic stress to psychological distress and sleep quality among adolescents (Kristensen et al., 2023; Yan et al., 2018). These findings suggest that academic stress involves multiple responses to academic demands and should therefore be assessed using instruments whose item content and measurement properties adequately represent the construct.

Previous studies have examined academic stress in relation to students' psychological well-being and behavior. Longitudinal associations between academic stress, academic self-efficacy, and psychological distress have been investigated among high school students to understand how these variables are related over time (Kristensen et al., 2023) . Among adolescents, academic stress has also been associated with sleep quality, with school burnout and depression examined as mediating variables in this relationship (Yan et al.,

2018) . In Indonesia, academic stress among adolescents has been measured using the *Educational Stress Scale for Adolescents* (ESSA) in a study examining its association with emotional eating behavior (Ramadhani & Mahmudiono, 2021) . A systematic review of 60 studies found that levels of distress among students in the final years of secondary school varied across countries and according to the measurement instruments used (Wuthrich et al., 2020) . The review also identified associations between distress and individual characteristics, lifestyle factors, and students' relationships with family members, schools, and peers (Wuthrich et al., 2020) . In addition to examining factors associated with academic stress, other research has investigated school-based programs designed to reduce or prevent academic stress among secondary school students (Wuthrich et al., 2020). These diverse research perspectives demonstrate that academic stress is a relevant construct in studies of education and adolescent well-being, highlighting the need to pay close attention to the quality of instruments used to measure it. Taken together, these studies demonstrate the broad application of academic stress measures across research contexts. However, findings about the consequences or correlates of academic stress do not, by themselves, establish that the instruments used measure the construct adequately in every population. Differences in the instruments and contexts of application make it important to examine how individual items and response categories function, as well as whether scores can be interpreted consistently for the intended population.

Measurement quality is fundamental to academic stress research because the conclusions drawn depend on how accurately an instrument represents the construct being measured. Instrument evaluation should examine item fit to the construct, the functioning of response categories, measurement reliability, and the instrument's ability to distinguish respondents according to their levels of the measured trait (Umar et al., 2025) . Differences in respondent characteristics or application contexts may also warrant an examination of whether items function consistently across groups (Umar et al., 2025). Specifically, the Rasch model has been used to evaluate an academic stress instrument among high school students in Bandung, including assessments of item fit, potential bias, and dimensionality (Yulianti et al., 2023) . This study provides an example of psychometric evaluation directly related to the measurement of academic stress (Yulianti et al., 2023) . However, findings from an evaluation conducted with one sample do not automatically guarantee that an instrument will demonstrate the same measurement properties in other populations or educational contexts. Therefore, the development and validation of an academic stress scale should include an examination of its measurement properties in relation to the intended use of the instrument and the characteristics of the respondents (Yulianti et al., 2023; Umar et al., 2025). The existence of this study means that the rationale for developing another scale cannot rest solely on the claim that Rasch analysis has not previously been applied to academic stress in this population. Instead, the contribution of the present study must be established through a direct comparison of the instruments and the evidence reported in the two studies. Relevant dimensions of comparison include the conceptual and item content of each instrument, the characteristics of the participants, the response-category structure, the item-fit and dimensionality results, the reliability and separation evidence, and the examination of differential item functioning (DIF). These comparisons are necessary to determine whether the present study extends previous work by evaluating different measurement features, using a distinct instrument or sample, or addressing an unresolved measurement question. The specific distinction between the present study and Yulianti et al. (2023) should therefore be stated in terms of documented differences rather than assumed limitations of the earlier study.

The Rasch model is a measurement approach that can be used to examine the relationship between respondents' response patterns and item characteristics within an instrument. In the validation process, Rasch analysis may include assessments of item fit, the effectiveness of response categories, person and item reliability and separation, and instrument dimensionality (Umar et al., 2025). Differential item functioning (DIF) analysis can be used to assess whether an item functions differently across groups after accounting for differences in the level of the measured construct (Umar et al., 2025). In academic stress measurement, Rasch analysis has been applied to evaluate item characteristics and instrument dimensionality among high school students (Yulianti et al., 2023) . The application of the same approach to an ego identity status instrument also illustrates its use in examining item fit, response category functioning, and reliability during instrument validation (Umar et al., 2025). Although the construct being measured differs, this study provides an example of how Rasch analysis can be applied to evaluate various aspects of measurement quality. These analyses address complementary measurement questions: whether items conform to the measurement model, whether response categories operate as intended, whether the instrument provides sufficiently consistent measurement, whether the items represent a sufficiently coherent dimension, and whether items function differently across groups after accounting for the measured trait. Consequently, instrument evaluation should be understood as an examination of specific measurement properties rather than as a single, general judgment that a scale is valid.

The present study addresses the measurement of academic stress through the development and Rasch evaluation of an instrument comprising items concerning students' emotional, cognitive, and behavioral reactions to academic demands. The initial instrument contained 22 items and was administered to 322 senior and vocational high school students in Indonesia. The analysis examined item fit, response category functioning, reliability, dimensionality, and gender-related DIF. This combination of analyses makes it possible to evaluate not only whether the retained items conform to the specified criteria but also whether the response scale functions adequately and whether some items may operate differently for female and male students. In this respect, the study provides an integrated examination of several measurement properties in the sample under investigation. Its contribution relative to Yulianti et al. (2023), however, should be interpreted in light of the precise instrument characteristics and analyses documented in that earlier report.

Based on this rationale, the present study aims to develop and validate an academic stress scale using Rasch analysis. Specifically, it examines item fit, response category functioning, measurement reliability and separation, dimensionality, and potential DIF, where the characteristics and size of the dataset permit such analyses. These evaluation areas are consistent with the use of the Rasch model to examine measurement quality during the validation of psychological instruments (Umar et al., 2025) and its application to academic stress measurement (Yulianti et al., 2023). The analysis focused on determining whether the items met the specified fit criteria, whether the response categories functioned appropriately, whether the scale demonstrated satisfactory reliability, and whether its dimensionality was consistent with the intended measurement structure. The DIF analysis was conducted to identify potential differences in item functioning between female and male students. The findings are expected to provide empirical evidence regarding the measurement properties of the academic stress scale and support further instrument refinement and subsequent evaluation in other student samples and educational contexts.

Method

This study employed a quantitative instrument validation design to examine the psychometric properties of an academic stress scale using the Rasch model. The analysis focused on response category functioning, item fit, person and item reliability and separation, dimensionality, and differential item functioning (DIF). The Rasch model was used to evaluate item performance and measurement consistency based on respondents' responses (Boone et al., 2013; Sumintono & Widhiarso, 2015).

The study involved 322 students aged 14–17 years who were enrolled in senior high school, comprising 190 female students and 132 male students. Participants were selected using a random sampling procedure. The inclusion criteria were students enrolled in senior high school who were between 14 and 17 years of age. The sample included students of both genders, allowing the examination of potential gender-related differences in item functioning through differential item functioning (DIF) analysis.

Table 1. Dimensions and Indicators of the Academic Stress Scale

Variable	Dimension	Indicators
Academic Stress	Emotional Reactions	Anxiety; irritability; panic; anger; dissatisfaction; unhappiness; restlessness
	Cognitive Reactions	Forgetfulness; difficulty concentrating; negative thinking; loss of self-confidence or feelings of worthlessness; persistent boredom or weariness, or lack of enjoyment in life; difficulty setting priorities
	Behavioral Reactions	Forgetfulness; difficulty concentrating; negative thinking; loss of self-confidence or feelings of worthlessness; persistent boredom or weariness, or lack of enjoyment in life; difficulty setting priorities

The three dimensions represent different types of responses associated with academic stress. Emotional reactions refer to students' emotional experiences, cognitive reactions involve thought processes and perceptions, and behavioral reactions encompass observable responses to academic demands. Table 1 presents the dimensions and indicators used to structure the academic stress scale.

Rasch analysis was conducted using Winsteps software to evaluate the psychometric properties of the instrument. Response category functioning was examined using observed average measures, category probability curves, and Andrich thresholds to assess whether the response categories functioned in an

ordered and distinguishable manner (Sumintono & Widhiarso, 2015) . Item fit was assessed using outfit mean-square (MNSQ), outfit standardized mean-square (ZSTD), and point-measure correlation statistics. The criteria used to evaluate item fit were 0.5 < outfit MNSQ < 1.5, −2.0 < outfit ZSTD < +2.0, and 0.4 < point-measure correlation < 0.85 (Boone et al., 2013; Sumintono & Widhiarso, 2015). These criteria were used to evaluate the consistency of item responses with the Rasch measurement model.

Person and item reliability and separation indices were examined to assess measurement consistency and the instrument's ability to distinguish among different levels of the measured trait. Dimensionality was evaluated using the percentage of raw variance explained by the measures and the unexplained variance in the first contrast (Sumintono & Widhiarso, 2015) . The numerical thresholds used to interpret these dimensionality statistics should be explicitly reported, together with their methodological references. DIF analysis was conducted to examine potential differences in item functioning between female and male students. A probability value below 0.05 was used as the criterion for identifying statistically significant DIF.

Results and Discussions

The response category functioning analysis revealed that the original five-category response scale did not fully meet the recommended threshold criteria. In the initial analysis, the threshold differences between Categories 1 and 2 and between Categories 3 and 4 were 1.54 and 1.76 logits, respectively, whereas the differences between Categories 2 and 3 and between Categories 4 and 5 were both 0.66 logits. According to Sumintono & Widhiarso (2015) , differences between adjacent Andrich thresholds should ideally range from 1.4 to 5.0 logits. Threshold differences below 1.4 logits may indicate that respondents do not clearly distinguish between adjacent response categories. Thus, the initial results suggested that some response options in the five-category format might not function as distinctly as intended. Although each category displayed a distinct peak in the category probability curves, this pattern alone did not resolve the concern raised by the threshold differences.

To address this issue, two revised four-category response structures were examined. In the first structure, the observed average measures increased across the categories, and the threshold differences between Categories 1 and 2 and between Categories 3 and 4 fell within the recommended range. However, the threshold difference between Categories 2 and 3 was only 0.56 logits. In the second structure, the observed average measures also increased, and the threshold differences between Categories 1 and 2 and between Categories 2 and 3 met the recommended range. Nevertheless, the difference between Categories 3 and 4 remained at 0.56 logits. Consequently, neither revised structure fully satisfied the recommended threshold criteria.

The first revised four-category structure was retained because its observed average measures and category measures showed an increasing pattern across the response categories. In the retained format, favorable items were scored as follows: *Very Not Suitable* = 1, *Not Suitable* = 2, *Suitable* = 3, and *Very Suitable* = 4. The final category functioning results are presented in Table 2.

Table 2. Summary of Category Functioning of the Academic Stress Scale

Response Category	Score	Observed Average	Andrich Threshold	Category Measure
Very Not Suitable	1	−1.91	None	−3.88
Not Suitable	2	−0.58	−2.76	−0.90
Suitable	3	0.53	1.10	1.40
Very Suitable	4	1.27	1.66	3.05

Source: Winsteps output, Summary of Category Structure.

As shown in Table 2, the observed average measures increased progressively from −1.91 for Category 1 to 1.27 for Category 4. The category measures followed the same general pattern, increasing from −3.88 to 3.05. These results indicate an orderly progression in the observed responses: higher response categories were associated with higher measured levels of the measured construct. However, the difference between the Andrich thresholds for Categories 3 and 4 was 0.56 logits (1.66 − 1.10), which remained below the

recommended minimum of 1.4 logits (Sumintono & Widhiarso, 2015). This finding indicates that the two highest categories may not be sufficiently differentiated in terms of their threshold locations, despite the progressive increase in observed average measures and category measures.

The findings illustrate why rating-scale evaluation should consider several indicators rather than relying on a single result. The increasing observed averages support the ordered progression of responses, whereas the small threshold difference identifies a potential limitation in the separation of adjacent categories. Sumintono & Widhiarso (2015) emphasize the use of observed average measures and Andrich thresholds in evaluating rating-scale functioning. In this study, the first revised four-category structure was selected based on the overall pattern of the observed averages and category measures, not because all threshold criteria had been satisfied. Therefore, the retained response format represents the preferred structure among those examined, while the threshold result for Categories 3 and 4 remains a limitation that should be considered when interpreting the scale's measurement properties. Further validation with a new sample could examine whether the same response-category pattern is reproduced.

Item fit analysis was conducted using the Rasch model to examine whether the 22 items of the academic stress scale functioned consistently with the measurement model. Three fit statistics were evaluated: outfit mean square (MNSQ), outfit Z-standard (ZSTD), and point-measure correlation (Pt Measure Corr). The acceptable ranges were $0.5 < \text{MNSQ} < 1.5$, $-2.0 < \text{ZSTD} < +2.0$, and $0.4 < \text{Pt Measure Corr} < 0.85$, respectively (Sumintono & Widhiarso, 2015) . These indicators provide complementary information about the extent to which item response patterns conform to the Rasch measurement model (Boone et al., 2013).

The item-fit output showed that several items did not meet one or more of the specified criteria. Item 6 had an outfit MNSQ of 1.67, exceeding the upper reference limit of 1.5, indicating a potential inconsistency between its observed response pattern and the model expectation. Items 6 and 1 also had outfit ZSTD values of 7.3 and 5.0, respectively, both exceeding the upper limit of +2.0. These results suggest that the responses to Items 6 and 1 deviated from the model expectations and therefore warrant further examination. Although Item 13 had an outfit ZSTD value of 1.9, this value remained within the specified range and should not be classified as misfitting on the basis of ZSTD alone.

The point-measure correlation results provided additional information about item functioning. Items 6, 1, 22, and 2 had correlation values of 0.21, 0.22, 0.30, and 0.25, respectively, all below the minimum criterion of 0.40. These low correlations suggest that responses to these items were less consistent with the overall measurement pattern than required by the study's criterion. In contrast, Item 13 had a point-measure correlation of 0.60, which fell within the acceptable range. Therefore, although Item 13 showed an outfit MNSQ of 1.20 and an outfit ZSTD of 1.9, its point-measure correlation did not indicate a problem under the specified criterion.

Table 3. Summary of Misfitting Items by Item Fit Order

Item Fit Criteria	Reference Range	Items Outside the Range
Outfit MNSQ	0.5–1.5	Item 6
Outfit ZSTD	−2.0 to +2.0	Items 6 and 1
Pt Measure Corr	0.4–0.85	Items 6, 1, 22, and 2
Items reported as misfitting in the study	—	Items 1, 2, 6, and 22
Items retained in the study	—	Items 3, 4, 5, and 7–21

Taken together, the results indicate that Items 1, 2, 6, and 22 were identified as misfitting in the study's overall classification, although the individual statistics did not show that all four items violated every criterion. Item 6 showed the clearest evidence of misfit across the three indicators, whereas Items 1, 2, and 22 showed low point-measure correlations. These differences highlight the importance of interpreting fit statistics individually and collectively rather than assuming that every misfitting item exhibits the same pattern.

Rasch fit statistics can help identify items whose response patterns do not adequately conform to the measurement model; however, statistical misfit does not necessarily mean that an item is theoretically irrelevant (Boone et al., 2013). The content and wording of each item should also be reviewed to determine

whether it adequately represents academic stress and its emotional, cognitive, or behavioral dimensions. Based on the study's overall classification, Items 1, 2, 6, and 22 were excluded, leaving 18 items for subsequent analyses. This decision should be interpreted in conjunction with the theoretical coverage of the retained items to ensure that the removal of misfitting items does not narrow the construct being measured.

The interpretation of item fit should distinguish statistical conformity from substantive importance. Rasch analysis identifies response patterns that are unexpected under the model, but a misfit statistic does not, by itself, explain why an item behaves unexpectedly or establish that its content is unimportant (Boone et al., 2014; Bond et al., 2020). The combination of outfit MNSQ, outfit ZSTD, and point-measure correlation is useful because the indicators describe different aspects of item response behavior. In the present results, Item 6 showed concerns across all three indicators, whereas Items 1, 2, and 22 were identified particularly through low point-measure correlations. These differing patterns suggest that the items should not be treated as psychometrically identical cases, even though they were grouped together in the study's overall exclusion decision. A nuanced interpretation preserves the distinction between the strength of the statistical evidence for each item and the final decision made for the scale.

Item fit is one source of evidence about whether the response data behave consistently with the proposed measurement model; it is not a substitute for theoretical justification or other forms of validity evidence (Aziz, 2011; Bond et al., 2020; Khine, 2020). The exclusion of four items reduced the scale from 22 to 18 items, but the shorter item set should not automatically be assumed to be superior simply because problematic items were removed. A shorter scale can be more efficient, yet it must still cover the construct sufficiently. Because academic stress in this study is conceptualized through emotional, cognitive, and behavioral reactions to academic demands, the retained statements should continue to represent these aspects. The item-fit results therefore provide a basis for refinement while making content review important in determining whether the revised scale remains conceptually comprehensive (Lazarus, 1999; Bond et al., 2020).

The item-fit findings also demonstrate why decisions should not rely on a single numerical threshold in isolation. Outfit MNSQ indicates the degree of unexpected response variation, ZSTD expresses the statistical extremity of that deviation, and point-measure correlation indicates how item responses align with the overall measure. Rasch analysis uses these indicators to evaluate item behavior in relation to the measurement construct (Boone et al., 2014; Bond et al., 2020). In this study, Item 6 had the clearest convergence of evidence across the three indicators, while Items 1, 2, and 22 had low point-measure correlations. The results identify items requiring attention, but the strength and type of evidence are not identical for every item. This distinction makes the item-removal decision more transparent and avoids overstating the degree to which each excluded item violated the model.

The reliability of the academic stress scale was evaluated using person reliability, item reliability, Cronbach's alpha, and separation indices. These statistics provide complementary information about the consistency of respondents' measures, the stability of the item hierarchy, and the internal consistency of responses. The results are summarized in Table 4.

Table 4. Summary of Reliability and Separation Statistics of the Academic Stress Scale

No.	Statistic	Value	Interpretation
1	Person reliability (real model)	0.79	Adequate
2	Person reliability (model)	0.83	Good
3	Item reliability (real model)	0.98	Excellent
4	Item reliability (model)	0.98	Excellent
5	Cronbach's alpha	0.84	Very good

Source: Rasch analysis output for 322 respondents and 18 items.

As shown in Table 4, Cronbach's alpha was 0.84, indicating good internal consistency according to the interpretation criteria reported by Sumintono and Widhiarso (2015). This result suggests that the responses to the 18 retained items demonstrated a satisfactory level of consistency within the study sample. However, Cronbach's alpha reflects overall internal consistency and should not be interpreted on its own as evidence that all items measure a single dimension.

Cronbach's alpha provides an overall indication of the consistency among item responses, whereas Rasch reliability statistics provide additional information about the reproducibility of respondent measures and item calibrations. Therefore, interpreting alpha alongside person and item reliability offers a broader evaluation of the instrument's measurement properties than relying on internal consistency alone (Boone et

al., 2014). Sari and Mahmudi (2024) also describe Cronbach's alpha and Rasch reliability indices as complementary information for evaluating assessment quality.

Person reliability was 0.79 in the real model and 0.83 in the model estimate. These values indicate adequate to good consistency in distinguishing respondents according to their measured levels of academic stress. The difference between the two estimates reflects the distinction between reliability calculated using observed error and that estimated under the model. The person separation indices were 1.95 and 2.24, respectively, suggesting that the scale could distinguish respondents across different levels of academic stress, although the degree of differentiation was not exceptionally high. In Rasch measurement, person reliability and separation are useful for evaluating how consistently and distinctly respondents can be differentiated by the instrument (Sumintono & Widhiarso, 2015) .

Person reliability and person separation should be interpreted together because they describe related but distinct aspects of measurement. Person reliability indicates the reproducibility of distinctions among respondents, while person separation reflects the instrument's ability to distinguish respondents at different locations on the measured continuum (Boone et al., 2014). Sari and Mahmudi (2024) similarly discuss separation indices as indicators of the extent to which respondents can be differentiated according to their measured characteristics.

In this study, the real-model person reliability of 0.79 and separation index of 1.95 suggest that respondent differentiation was adequate but not particularly strong. The model estimates of 0.83 and 2.24 provide somewhat stronger evidence of differentiation. These findings suggest that the scale can identify differences in students' academic stress levels.

Item reliability was 0.98 for both the real and model estimates, while item separation was 7.52 and 7.85, respectively. These high values indicate that the relative ordering of item difficulty was highly stable within the analyzed sample and that the items were distributed across a broad range of the measured continuum. Boone et al. (2013) explain that Rasch reliability and separation statistics provide information about the reproducibility of item calibrations and the capacity of a measurement instrument to distinguish different locations along the measurement continuum. In this study, the high item reliability and separation values therefore support the stability of the item hierarchy in the analyzed sample.

The results suggest that the academic stress scale demonstrated good internal consistency and strong item reliability, while person reliability indicated adequate-to-good consistency in differentiating respondents. The comparatively higher item reliability than person reliability suggests that the item hierarchy was more stable than the distinctions among respondents' measured stress levels. Nevertheless, these findings should be interpreted alongside the item-fit and dimensionality results because reliability alone does not establish the validity of the scale. Further validation with other student samples would help determine whether the reliability estimates and item hierarchy remain stable across populations.

Differential item functioning (DIF) analysis was conducted to examine whether items on the academic stress scale functioned differently for male and female respondents. An item was considered to show statistically significant DIF when its probability value was below .05. DIF analysis is important in instrument validation because items that function differently across groups may affect the interpretation and fairness of comparisons between respondents (Boone et al., 2013).

The analysis identified eight items with statistically significant DIF: Items 4, 7, 8, 9, 18, 19, 20, and 21. Their probability values were .0437, .0000, .0033, .0004, .0198, .0030, .0002, and .0002, respectively. The remaining ten items did not meet the significance criterion. These results indicate that the response patterns for the eight identified items differed statistically between the gender groups under the DIF analysis.

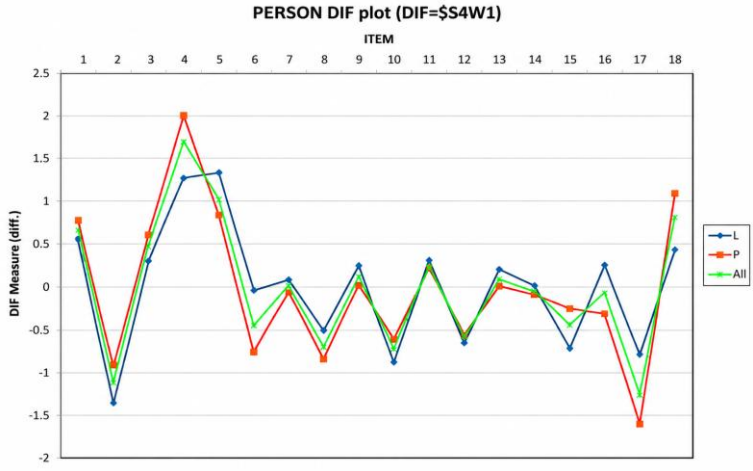


Figure 2. Person DIF Plot by Gender

The initial interpretation of the DIF plot indicated that female respondents were relatively favored on Items 2, 4, 15, and 18, whereas male respondents were relatively favored on Items 5, 6, 16, and 17. This pattern suggests that the two gender groups may have interpreted or responded to these statements differently, even after accounting for their measured levels of academic stress. Such differences do not necessarily indicate that one gender experiences greater academic stress overall. Rather, they may reflect differences in how particular items function across groups.

Table 4. Summary of Items Showing Differential Item Functioning by Gender

Item Number (Probability)	Gender Relatively Favored	Interpretation
Item 2 ($p < .05$)	Female	Significant DIF; relatively favors female respondents
Item 4 ($p < .05$)	Female	Significant DIF; relatively favors female respondents
Item 5 ($p < .05$)	Male	Significant DIF; relatively favors male respondents
Item 6 ($p < .05$)	Male	Significant DIF; relatively favors male respondents
Item 15 ($p < .05$)	Female	Significant DIF; relatively favors female respondents
Item 16 ($p < .05$)	Male	Significant DIF; relatively favors male respondents
Item 17 ($p < .05$)	Male	Significant DIF; relatively favors male respondents
Item 18 ($p < .05$)	Female	Significant DIF; relatively favors female respondents

The findings highlight the need to examine the content of the eight flagged items more closely. In Rasch measurement, DIF can indicate that an item behaves differently across groups and therefore warrants further investigation (Boone et al., 2013). For the academic stress scale, the wording and content of each flagged statement should be reviewed to determine whether it may be interpreted differently by male and female students. Items should not automatically be removed solely because their DIF probability is statistically significant. Decisions should also consider their theoretical relevance and contribution to the emotional, cognitive, and behavioral dimensions of academic stress.

The initial DIF interpretation provides preliminary evidence that some items may function differently by gender. Further validation should establish whether these differences are reproducible in an independent sample and whether revisions are necessary to support fairer comparisons between male and female respondents. The final interpretation should be based on a consistent set of item-level probability values and DIF contrasts.

The response-category results should also be considered in relation to the purpose of the response format. A rating scale is useful when respondents can map their experiences onto response options that are meaningfully distinct, rather than selecting adjacent labels that feel similar. Linacre (2007) emphasizes that rating-scale quality depends on ordered category functioning and adequate distinctions between category thresholds. Kulas et al. (2008) also examined the functioning of middle response options in Likert-type responses, showing that the way response options are designed and interpreted is relevant to response behavior. This is relevant to academic stress because students may experience stress along a continuum and may find it difficult to distinguish adjacent options unless the labels communicate clear differences. The findings therefore support evaluating both the statistical behavior and wording of the response options rather than assuming that reducing the number of categories automatically improves measurement quality (Sumintono & Widhiarso, 2015; Linacre, 2007; Kulas et al., 2008).

From a measurement-development perspective, the choice between a five-category and a four-category format should be justified by evidence that respondents use the categories consistently with the intended interpretation of the scale. The selected four-category structure showed ordered observed averages and category measures, but the 0.56-logit distance between the two upper thresholds remained below the recommended range. This is a mixed result: the category sequence was ordered, while the distinction

between some adjacent categories may still be limited. Therefore, the retained format is best described as the preferred option among those examined, rather than as a fully optimal response structure. This interpretation is consistent with the principle that rating-scale quality should be judged through several indicators, including category ordering and threshold behavior (Linacre, 2007; Sumintono & Widhiarso, 2015).

Conclusions

This study provides preliminary evidence regarding the measurement properties of the Academic Stress Scale based on Rasch analysis. The 18 retained items demonstrated good internal consistency and high item reliability, while the dimensionality analysis met the criteria applied in this study. However, the four-category response structure did not fully satisfy the recommended threshold criteria, and several items exhibited gender-related differential item functioning (DIF). These findings indicate that, although some measurement properties were supported by the results, further refinement and evaluation are needed to strengthen the measurement properties of the scale. The findings also suggest that the interpretation and comparability of scores between female and male students require caution because gender-related DIF may affect how some items function across groups. Therefore, further refinement and validation are needed before the scale can be confidently applied across different groups of senior high school students. Future studies should examine response category functioning, investigate the magnitude and implications of gender-related DIF, and evaluate the scale in more diverse student samples and educational contexts. Additional evidence is needed to establish the scale's measurement properties and support its application across different student groups.

Acknowledgments

The authors would like to express their sincere gratitude to everyone who contributed to the preparation of this article. The authors greatly appreciate the assistance, collaboration, feedback, and support provided throughout the research process and manuscript preparation. Their contributions were valuable to the completion of this study.

References

Aziz, A. A. (2011). *Rasch model fundamentals: Scale construct and measurement structure*. Malaysia: Integrated Advance Planning Sdn Bhd.

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). New York, NY: Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-007-6857-4

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095. https://www.rasch.org/rmt/rmt211m.htm

Jagiello, T., Belcher, J., Neelakandan, A., Boyd, K., & Wuthrich, V. M. (2025). Academic stress interventions in high schools: A systematic literature review. *Child Psychiatry & Human Development, 56*, 1836–1869. https://doi.org/10.1007/s10578-024-01667-5

Khine, M. S. (Ed.). (2020). *Rasch measurement: Applications in quantitative educational research*. Singapore: Springer. https://doi.org/10.1007/978-981-15-1800-3

Kristensen, S. M., Larsen, T. M. B., Urke, H. B., & Danielsen, A. G. (2023). Academic Stress, Academic Self-efficacy, and Psychological Distress: A Moderated Mediation of Within-person Effects. *Journal of Youth and Adolescence*, *52*(7), 1512–1529. https://doi.org/10.1007/s10964-023-01770-1

Kristensen, S. M., Larsen, T. M. B., Urke, H. B., & Danielsen, A. G. (2023). Academic stress, academic self-efficacy, and psychological distress: A moderated mediation of within-person effects. *Journal of Youth and Adolescence, 52*(7), 1512–1529. https://doi.org/10.1007/s10964-023-01770-1

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22*(3), 251–259. https://doi.org/10.1007/s10869-008-9064-2

Lazarus, R. S. (1999). *Stres and Emotion A New Synthesis*. Springer Publishing Company.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. New York, NY: Springer Publishing Company.

Linacre, J. M. (2007). *[Referensi perlu diverifikasi: judul "Rating Scale Instrument Quality Criteria" tercatat atas nama W. P. Fisher, Jr.]*

Ramadhani, N., & Mahmudiono, T. (2021). Academic stress is associated with emotional eating behavior among adolescent. *Media Gizi Indonesia, 16*(1), 38–47. https://doi.org/10.20473/mgi.v16i1.38-47

Ramadhani, N., & Mahmudiono, T. (2021). Attribution-NonCommercial-ShareAlike license (CC BY-NC-SA 4.0). ACADEMIC STRESS IS ASSOCIATED WITH EMOTIONAL EATING BEHAVIOR AMONG ADOLESCENT. *Media Gizi Indonesia (National Nutrition Journal). 2021*, *16*(1), 38–47. https://doi.org/10.204736/mgi.v16i1

Sari, E. D. K., & Mahmudi, I. (2024). *Analisis pemodelan Rasch pada assessment pendidikan*. PT Pena Persada Kerta Utama.

Sumintono, B. , & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan*. Trim Komunikata.

Umar, N. F., Syahril, M. F., Nasution, S., Ardis, N., Azzahrah, H., & Rafli, M. (2025). Validity and reliability of the Objective Measure of Ego Identity Status adaptation: The Rasch Model approach. *Bulletin of Counseling and Psychotherapy, 7*(3). https://doi.org/10.51214/002025071583000

Wuthrich, V. M., Jagiello, T., & Azzi, V. (2020). Academic stress in the final years of school: A systematic literature review. *Child Psychiatry & Human Development, 51*(6), 986–1015. https://doi.org/10.1007/s10578-020-00981-y

Yan, Y. W., Lin, R. M., Su, Y. K., & Liu, M. Y. (2018). The relationship between adolescent academic stress and sleep quality: A multiple mediation model. *Social Behavior and Personality*, *46*(1), 63–77. https://doi.org/10.2224/sbp.6530

Yulianti, F., & Saripah, I. (2023). Harapan siswa yang mengalami stres akademik. *G-Couns: Jurnal Bimbingan dan Konseling, 8*(1), 413–420. https://doi.org/10.31316/gcouns.v8i01.5261

Yusuf, N. M., & Yusuf, J. M. (2020). *Faktor-Faktor Yang Mempengaruhi Stres Akademik*. *13*(2). 235–239. https://doi.org/10.35134/jpsy165.v13i2.84