



Character assessment reimagined: an IRT-based graded response model for honesty and discipline

Yakob Luturmas*, Kristina Luturdas

STPAK Santo Yohanes Penginjil, Ambon, Maluku, Indonesia

*Correspondence author: yakobluturmas@stpakambon.ac.id

DOI: <https://doi.org/10.65881/integration.v1i2.181>

ARTICLE INFO

History:

Received: 08-13-2026

Revised: 08-18-2026

Accepted: 08-19-2026

Published: 08-25-2026

Keywords:

honesty;
discipline;
character assessment;
item response theory;
graded response model.

ABSTRAK

Purpose: to develop and validate an IRT-based graded response model (GRM) instrument for assessing honesty and discipline among junior secondary school students.

Method: This study employed a research and development (R&D) approach involving 226 junior secondary school students. A 30-item likert instrument was evaluated using Aiken's V, confirmatory factor analysis (CFA), and the graded response model (GRM) to examine its validity, item characteristics, model fit, and measurement precision.

Findings: the developed instrument demonstrated strong content and construct validity, high internal consistency, and good item functioning. The GRM analysis indicated that the items had strong discrimination and satisfactory model fit, while the test information function showed adequate measurement precision, particularly across the lower-to-middle levels of honesty and discipline. Overall, the instrument is suitable for diagnostic and early assessment of students' honesty and discipline.

Implications: the study provides a psychometrically sound and evidence-based instrument for assessing students' honesty and discipline, supporting more objective character assessment and data-informed character education practices in junior secondary schools.

Originality: lies in the integration of character-based indicators of honesty and discipline with an Item response theory-graded response model (IRT-GRM) to develop a more precise and objective affective assessment instrument for junior secondary school students.



Introduction

Education is not solely concerned with the development of cognitive competencies but also with the formation of affective dispositions that shape students' behavior, responsibility, and moral conduct (Hasriani et al., 2025). In the context of twenty-first-century education, character education has become increasingly important because schools are expected not only to develop academically competent learners but

also to foster individuals who demonstrate integrity, responsibility, and strong moral values (Rosita et al., 2022; Tan, 2025). Among the fundamental character values required in both academic and social contexts are honesty and discipline (Ramadhani et al., 2024). Honesty reflects consistency between what an individual says and does and serves as a foundation for trust, integrity, and ethical behavior (Cooper et al., 2023). Discipline, meanwhile, refers to an individual's willingness to comply with rules, manage behavior, and consistently fulfill responsibilities (Gustiawan et al., 2025). These two characteristics are particularly important during adolescence, including the junior secondary school period, when students experience substantial developmental and behavioral transitions that may influence the formation of long-term patterns of conduct (Carrera et al., 2026; Uge et al., 2022).

Despite their importance, the assessment of students' character and affective dispositions remains a challenging issue in educational practice (Bowers & Sabin, 2024). Honesty and discipline are psychological constructs that cannot be directly observed or measured in the same manner as cognitive outcomes (Paap et al., 2023). Instead, they represent latent traits that must be inferred from observable indicators and patterns of responses to carefully designed assessment items (Reise & Moore, 2023; Stefana et al., 2025). In schools, however, affective assessment is still frequently conducted through teacher observations and informal judgments without sufficiently standardized measurement instruments (Monita & Hasan, 2023). Such practices may introduce subjectivity and inconsistency, potentially resulting in inaccurate representations of students' actual character profiles (Ahmad et al., 2024; Ashwin et al., 2026). This problem is particularly consequential when assessment results are expected to support character development, because weak measurement can lead to inappropriate identification of students' strengths and weaknesses and limit the usefulness of assessment information for instructional decision-making.

A further methodological concern arises from the predominance of classical test theory (CTT) in the development and evaluation of affective instruments. Although CTT has made an important contribution to educational and psychological measurement, its item and test statistics are substantially dependent on the characteristics of the sample used for calibration. Moreover, CTT generally provides less detailed information about how individual items function across different levels of the underlying trait (Baker, 2001; Reise & Moore, 2023). For character assessment, this limitation is important because an item may function effectively for students with moderate levels of honesty or discipline but provide little information for students at very low or very high levels of the trait (Jeynes, 2019). Consequently, an assessment instrument should not only demonstrate overall reliability and validity but should also provide evidence that individual items appropriately discriminate among students across the latent-trait continuum (Gempp & Carrasco, 2025).

Item response theory (IRT) provides a stronger theoretical and methodological framework for addressing these limitations. IRT models the probability of a particular response as a function of an individual's position on a latent trait and the characteristics of the item (Baker, 2001; Reise & Moore, 2023). Unlike conventional CTT-based analysis, IRT enables researchers to examine item-level characteristics, including discrimination and the location of response thresholds, while providing information about how much measurement information an item contributes at different levels of the latent construct (Gyamfi & Acquaye, 2023). This makes IRT particularly relevant for the assessment of affective characteristics, where the objective is to obtain a more precise estimate of an underlying disposition rather than merely calculate a total observed score (Zitzmann et

al., 2024). In addition, contemporary applications of IRT demonstrate its usefulness for evaluating item functioning and measurement quality in complex assessment contexts (Belzak, 2023).

For instruments using ordered likert-type response categories, the graded response model (GRM) proposed by Samejima (1969) is particularly appropriate. The GRM is designed for polytomous items in which response categories represent increasing levels of an underlying trait (Mansour et al., 2026). Rather than treating a likert response as a simple numerical score, the model estimates the probability that a respondent will select a category at or above each response threshold (Carolis et al., 2026; Jin et al., 2022). This characteristic allows researchers to examine whether the response categories meaningfully distinguish different levels of honesty and discipline and to identify the regions of the latent continuum in which individual items provide the greatest amount of information. The use of GRM therefore offers a theoretically coherent approach for transforming ordinal responses into more informative estimates of students' latent character traits (Reise & Moore, 2023; Samejima, 1969). From a practical perspective, IRT analysis can also be integrated with established content-validation procedures, such as Aiken's V, to ensure that items are substantively relevant before their psychometric functioning is examined empirically (Ahmad et al., 2024; Contreras et al., 2025).

The existing literature provides evidence supporting the importance of systematic character assessment and the use of more rigorous measurement approaches. Research has highlighted the importance of discipline and responsibility as components of character development in educational settings (Rosita et al., 2022; Saputra et al., 2024; Uge et al., 2022). At the methodological level, previous studies have demonstrated the usefulness of Aiken's V for evaluating content validity and IRT for obtaining more detailed information about item functioning (Ahmad et al., 2024; Contreras et al., 2025). Nevertheless, these strands of research have not yet been sufficiently integrated into the development of a dedicated instrument for measuring honesty and discipline among junior secondary school students. Moreover, the available literature remains limited in applying the GRM specifically to the simultaneous assessment of these two character dimensions within the Indonesian junior secondary school context. Existing character studies tend to emphasize the implementation or cultivation of discipline and responsibility rather than the development of a rigorously calibrated measurement instrument (Saputra et al., 2024; Uge et al., 2022). This creates a research gap concerning how honesty and discipline can be operationalized as latent traits and measured through an empirically validated, item-level IRT framework.

The present study addresses this gap by developing and validating an IRT-based graded response instrument for assessing honesty and discipline among junior secondary school students. The novelty of the study lies not merely in producing another character questionnaire, but in integrating substantive character indicators with systematic content validation and GRM-based psychometric analysis. This approach is expected to provide evidence regarding the functioning of individual items, their discrimination characteristics, response-category thresholds, and the amount of information they provide across different levels of honesty and discipline. Such an approach moves character assessment beyond a simple aggregation of likert scores toward a more sophisticated measurement framework in which students' responses are interpreted in relation to underlying latent traits.

Accordingly, the primary objective of this study is to develop and evaluate the psychometric quality of an instrument for measuring students' honesty and discipline

using the graded response model. Specifically, the study seeks to establish the content validity of the instrument, examine the functioning and quality of its items through IRT, and determine whether the resulting instrument can provide sufficiently precise measurement across the relevant levels of the latent character traits. The theoretical foundation of the study is therefore derived from the conceptualization of honesty and discipline as latent affective constructs, the principles of IRT as a framework for modeling latent traits through item responses, and Samejima's GRM as an appropriate model for ordered polytomous responses (Baker, 2001; Reise & Moore, 2023; Samejima, 1969).

The importance of conducting this research is both methodological and educational. Methodologically, a well-calibrated GRM-based instrument can address several weaknesses associated with subjective and predominantly CTT-based character assessment by providing more detailed evidence at the item and latent-trait levels. Educationally, a valid and precise instrument can provide teachers and schools with more dependable information about students' character-related dispositions. Such information can support the identification of areas requiring reinforcement, the evaluation of character education programs, and more evidence-based decision-making in student development. In this sense, improving the quality of character measurement is not an end in itself; rather, it is a prerequisite for ensuring that interventions designed to strengthen honesty and discipline are informed by credible assessment evidence.

Ultimately, this study is expected to contribute to educational measurement, character education, and school-based assessment in several ways. First, it is expected to provide a theoretically grounded and empirically validated instrument for assessing honesty and discipline in the junior secondary school context. Second, it will contribute methodological evidence regarding the application of the GRM to affective and character assessment, particularly for ordered likert-type responses. Third, the study may provide practitioners with a more objective basis for understanding students' character-related dispositions and supporting targeted character-development strategies. Finally, the study is expected to contribute to the broader development of data-informed character education in Indonesia, in which assessment is not limited to subjective judgment but is supported by rigorous psychometric evidence.

Method

This study employed a research and development (R&D) approach aimed at developing and validating a psychometrically sound instrument for assessing honesty and discipline among junior secondary school students. The development procedure followed a systematic sequence consisting of needs analysis, construct identification, test blueprint development, item construction, expert validation, readability testing, field testing, data analysis, and instrument revision (Mardapi, 2008). This procedure was designed to ensure that the developed instrument was theoretically relevant, empirically supported, and capable of providing accurate and consistent measurements of students' honesty and discipline.

The participants consisted of 226 junior secondary school students from several schools in Ambon city, Indonesia. Participants were selected using a simple random sampling technique. The sample size was considered adequate for the application of item response theory (IRT) because an appropriate number of respondents is required to obtain relatively stable estimates of item parameters (Contreras et al., 2025). The study focused on two target constructs, namely honesty and discipline, which were

operationalized through observable behavioral indicators derived from the conceptual framework and relevant literature.

The instrument development process began with the identification of dimensions and behavioral indicators for honesty and discipline, followed by the preparation of an instrument blueprint. Based on the blueprint, 30 Likert-type items were developed to represent the two constructs. Each item used four ordered response categories: always, often, rarely, and never. The instrument was administered to participants in digital form using Google Forms. Before the field trial, the items were subjected to expert content validation and readability evaluation. Content validity was assessed using Aiken's *V* coefficient, involving three experts with expertise in psychology, educational measurement, and educational practice. The experts evaluated the relevance, clarity, and appropriateness of each item in representing the intended construct. Items that did not meet the predetermined criteria were revised before the instrument was administered in the field.

The collected data were analyzed using R statistical software with the *mirt* package (Chalmers, 2012). The analysis was conducted in several stages. First, the dimensionality of the instrument was examined through exploratory factor analysis (EFA) to investigate the empirical structure underlying the item responses. Subsequently, confirmatory factor analysis (CFA) was conducted to evaluate the extent to which the hypothesized measurement structure was supported by the empirical data. Appropriate model-fit indices were considered to determine the adequacy of the proposed factor structure and to provide evidence for the construct validity of the instrument.

Following the dimensionality analysis, the instrument was analyzed using the graded response model (GRM), an IRT model designed for ordered polytomous response categories. The GRM was used to estimate the principal item parameters, particularly the discrimination parameter (*a*) and threshold parameters (*b*). The discrimination parameter indicates the extent to which an item can differentiate respondents with different levels of the underlying trait, whereas the threshold parameters indicate the locations on the latent-trait continuum at which respondents are increasingly likely to select higher response categories. Examination of these parameters provided detailed information about how each item functioned across different levels of honesty and discipline.

The fit of the GRM was evaluated using the $S-X^2$ item-fit statistic to determine whether the observed response patterns were consistent with the model expectations. Items exhibiting inadequate model fit were examined together with their content validity and estimated item parameters to determine whether they required revision or removal. In addition, measurement precision was evaluated using the test information function (TIF) and standard error of measurement (SEM). The TIF was used to determine the amount of information provided by the instrument at different levels of the latent traits, while the SEM was used to identify the degree of measurement error across the latent-trait continuum. These analyses enabled the researchers to determine the levels of honesty and discipline at which the instrument provided the most precise measurement.

Finally, evidence obtained from the content-validation process, dimensionality analysis, GRM parameter estimation, item-fit analysis, and measurement-precision analysis was integrated to determine the overall quality of the instrument. Items with insufficient content validity, weak discrimination, inappropriate threshold characteristics, or poor model fit were considered for revision or elimination based on

both statistical evidence and theoretical considerations. The final instrument was established by considering evidence of content validity, construct validity, item functioning, model fit, and measurement precision. Through this systematic R&D procedure, the study sought to produce an IRT-based instrument that provides valid, reliable, and sufficiently precise measurements of honesty and discipline among junior secondary school students.

Results and discussion

Initial product development

The initial development stage produced an affective assessment instrument designed to measure two character constructs among junior secondary school students: honesty and discipline. The development of the instrument was based on a needs analysis and a theoretical review of the two constructs. Honesty was operationalized as behavior that reflects consistency between statements and actual conditions without manipulation, whereas discipline was defined as behavior characterized by compliance with rules, punctuality, and consistency in fulfilling responsibilities. These operational definitions were subsequently translated into specific behavioral indicators to ensure that each item represented an observable aspect of the intended latent construct.

A total of 30 items were initially developed, consisting of 15 items for honesty and 15 items for discipline. Each item used four ordered response categories: always (SL), often (SR), rarely (JR), and never (TP), which were scored from 4 to 1, respectively. The distribution of the items across the two constructs and their behavioral indicators is presented in Table 1.

Table 1 instrument blueprint

Variable	Operational definition	Indicator	Item numbers
Honesty	Behavior of speaking and acting according to reality without manipulation	Telling things as they are	1–3
		Not lying	4–6
		Honesty in assignments/tests	7–9
		Admitting mistakes	10–12
		Not manipulating information	13–15
Discipline	Behavior characterized by rule compliance, punctuality, and consistency in fulfilling obligations	Compliance with rules	16–18
		Punctuality	19–21
		Consistency in learning	22–24
		Responsibility for assignments	25–27
		Orderliness in activities	28–30

As shown in Table 1, the instrument was designed to represent the two constructs in a balanced manner. Each construct was represented by five behavioral indicators, with three items assigned to each indicator. This distribution provided comparable representation across the honesty and discipline dimensions and established the basis for subsequent content validation and empirical testing.

Content validity

Content validity was examined through expert judgment involving three validators with expertise in psychology, educational measurement, and educational practice. The validators evaluated all 30 items with respect to their material relevance, construct representation, construction, and language using a four-point rating scale. A score of 1 indicated that an item was not relevant to the intended indicator, whereas a

score of 4 indicated that the item was highly relevant. In addition to assigning numerical ratings, the validators provided qualitative suggestions concerning the organization of indicators and the wording of individual statements. The qualitative evaluation resulted in several recommendations, particularly concerning the systematic ordering of indicators and the selection of verbs that more accurately reflected the behavioral indicators being measured. These recommendations were incorporated into the revision process by refining the wording, contextual appropriateness, and clarity of the items before the field trial. The numerical ratings obtained from the three validators were analyzed using Aiken's V coefficient. The results are summarized in Table 2.

Table 2 Aiken's V content validity analysis

Aspect	Number of items	Mean V	Minimum	Maximum	Category
Honesty and discipline	30	0.84	0.67	1.00	Very valid

Source: primary data, processed

As presented in Table 2, the 30 items obtained a mean Aiken's V coefficient of 0.84, with individual coefficients ranging from 0.67 to 1.00. The results indicate that the overall content validity of the instrument was high. The lowest coefficient was 0.67, while several items achieved the maximum coefficient of 1.00. Based on the predetermined validity criteria, none of the 30 items was eliminated solely on the basis of the content-validation results. The qualitative recommendations from the validators were therefore used primarily to improve item formulation and clarity. Before the field trial, a readability test was conducted with 10 junior secondary school students. The readability assessment indicated that the statements could be understood by the target respondents and did not produce substantial ambiguity. Based on the readability results, no further substantive revision was required before the instrument was administered in the main field test.

Construct validity

The empirical evaluation involved 226 junior secondary school students from several schools in Ambon City. Prior to factor analysis, the adequacy of the data for factor analysis was examined using the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity. The KMO value was 0.887, exceeding the minimum criterion of 0.50, while Bartlett's test produced a significance value below 0.05. These results indicated that the correlation matrix was suitable for factor analysis. Confirmatory factor analysis (CFA) was subsequently conducted to examine the hypothesized two-factor structure consisting of honesty and discipline. The goodness-of-fit results are presented in Table 3.

Table 3 goodness-of-fit indices for the confirmatory factor analysis

Fit index	Obtained value	Criterion	Conclusion
CFI	0.948	> 0.90	Fit
TLI	0.932	> 0.90	Fit
RMSEA	0.058	< 0.08	Fit
SRMR	0.049	< 0.08	Fit

Source: primary data, processed

As shown in Table 3, the CFA model demonstrated acceptable-to-good fit across all reported indices. The CFI was 0.948, and the TLI was 0.932, both exceeding 0.90. The RMSEA was 0.058, while the SRMR was 0.049, both below the 0.08 criterion.

Collectively, these results indicate that the hypothesized two-factor measurement model was adequately supported by the observed data.

Reliability

The internal consistency of the 30-item instrument was examined using Cronbach's alpha. The reliability results are presented in Table 4.

Table 4 reliability statistics

Statistic	Value
Cronbach's alpha	0.938
Cronbach's alpha based on standardized items	0.941
Number of items	30

Source: primary data, processed

As presented in Table 4, the instrument obtained a Cronbach's alpha coefficient of 0.938, while the standardized-item alpha was 0.941 across the 30 items. The corresponding standard error of measurement (SEM) for the total observed score was 2.98. These results indicate a high level of internal consistency and relatively small measurement error at the observed-score level.

Item characteristics based on the graded response model

The item-level characteristics were subsequently examined using the graded response model (GRM). The estimated parameters included the discrimination parameter (a) and threshold parameters (b). Selected item-parameter estimates are presented in Table 5.

Table 5 selected item parameters based on the graded response model

Item	Discrimination (a)	Threshold ($b1$)	Threshold ($b2$)	Description
B1	2.45	-2.12	-0.85	Very high discrimination
B10	1.88	-1.50	0.20	High discrimination
B16	2.10	-1.90	-0.45	High discrimination
B30	1.75	-1.25	0.55	High discrimination

Source: primary data, processed

The reported items showed discrimination parameters ranging from 1.75 to 2.45. B1 had the highest discrimination parameter ($a = 2.45$), followed by B16 ($a = 2.10$), B10 ($a = 1.88$), and B30 ($a = 1.75$). The threshold estimates shown in the table ranged from -2.12 to 0.55. B1 had the lowest first threshold ($b1 = -2.12$), whereas B30 had the highest second threshold ($b2 = 0.55$). According to the reported S-X² analysis, all items demonstrated acceptable model fit, with $p > 0.05$, indicating no statistically significant evidence of item misfit to the GRM.

Measurement precision: test information function and standard error of measurement

The measurement precision of the instrument across the latent-trait continuum was examined using the test information function (TIF) and standard error of measurement (SEM). The resulting curves are presented in Figure 1. As illustrated in Figure 1, the TIF curve showed relatively high information across a latent-trait range of approximately $\theta = -2.5$ to $+1.0$, with the highest concentration of information occurring within this interval. In the same region, the SEM curve reached relatively low values, indicating greater measurement precision. Outside this range, particularly at higher

levels of θ , the amount of information decreased, and the corresponding measurement error increased.

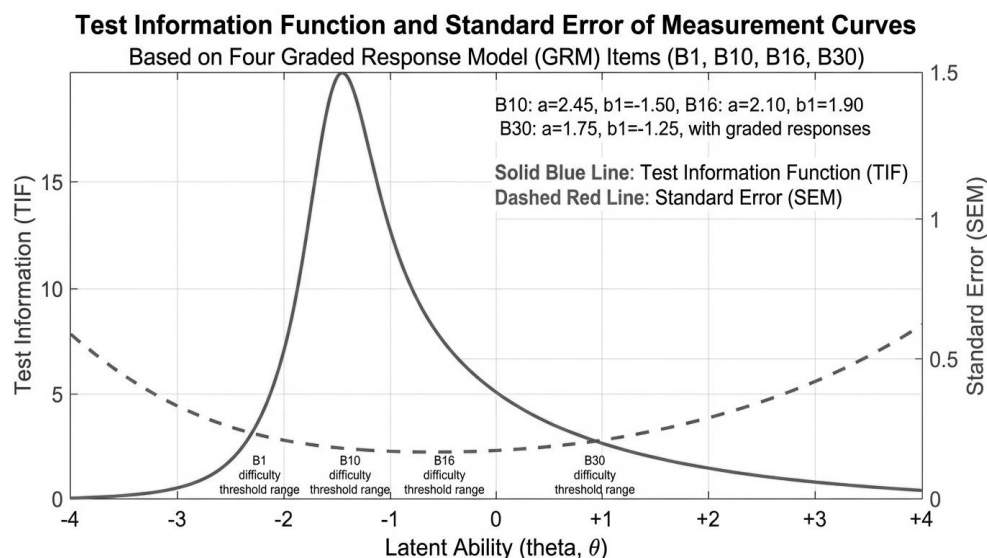


Figure 1 test information function and standard error of measurement

Source: primary data, processed

The findings demonstrate that the developed instrument possesses a coherent combination of content validity, construct validity, internal consistency, item discrimination, model fit, and measurement precision. More importantly, the results indicate that the development of an honesty and discipline instrument can be strengthened by integrating conventional validity procedures with an IRT-based analysis. This is particularly relevant because honesty and discipline are latent affective characteristics that cannot be observed directly and therefore require systematic inference from observable behavioral responses (Reise & Moore, 2023).

The first important finding concerns the content validity of the instrument. The mean Aiken's V coefficient of 0.84 indicates that the three experts generally judged the items to be relevant to the intended constructs. The range from 0.67 to 1.00 also shows that the degree of expert agreement varied across items. This finding is consistent with previous instrument-development studies that have employed expert judgment and Aiken's V to establish evidence of content validity (Ahmad et al., 2024; Contreras et al., 2025). Ahmad et al. (2024), for example, demonstrated the usefulness of Aiken's V for systematically evaluating the relevance of instrument items before empirical testing, while Contreras et al. (2025) similarly integrated expert-based content validation with subsequent psychometric evaluation. The present findings extend this approach to character assessment by applying expert validation to behavioral indicators of honesty and discipline.

The qualitative feedback from the validators is also important because content validity should not be interpreted solely through a numerical coefficient. The recommendation to improve the ordering of indicators and to align verbs more precisely with the intended behavioral indicators suggests that item clarity and behavioral specificity are critical in affective measurement. This is especially relevant for honesty and discipline because overly general statements may invite different interpretations among respondents. By contrast, statements referring to concrete and recognizable behaviors are more readily connected to the intended construct. Thus, the expert-review

process contributed not only statistical evidence of content relevance but also substantive refinement of the instrument.

The CFA results provide further support for the proposed measurement structure. The values of CFI = 0.948, TLI = 0.932, RMSEA = 0.058, and SRMR = 0.049 collectively indicate that the hypothesized two-factor model provides an adequate representation of the observed response data. These findings are consistent with contemporary recommendations that model fit should be evaluated using multiple complementary indices rather than relying on a single statistic (Sathyanarayana & Mohanasundaram, 2024). The consistency across incremental and residual-based indices strengthens the evidence that the proposed measurement structure is empirically plausible.

Substantively, the CFA findings support the conceptual distinction between honesty and discipline as related but distinguishable character dimensions. This distinction is important because both characteristics may be grouped broadly under the concept of character, yet they represent different behavioral tendencies. Honesty primarily concerns truthfulness, transparency, and the absence of deceptive behavior, whereas discipline emphasizes rule compliance, punctuality, consistency, and responsibility (Patri et al., 2025). The empirical support for a two-factor structure therefore provides a stronger basis for interpreting the instrument through separate honesty and discipline dimensions rather than automatically reducing all responses to a single general character score. Nevertheless, formal claims of discriminant validity would be strengthened by reporting additional evidence, such as the correlation between the latent factors and specific discriminant-validity statistics.

The high internal-consistency coefficient provides additional evidence regarding the reliability of the instrument. The Cronbach's alpha of 0.938 is substantially higher than the commonly used minimum criterion of 0.70, indicating strong consistency among the 30 items. This finding is broadly consistent with the validation approach used by Contreras et al. (2025), in which reliability evidence was considered together with structural validity to establish the psychometric quality of an instrument. The present study similarly demonstrates that the items operate consistently as indicators of the intended constructs.

However, the very high alpha coefficient should be interpreted carefully. A coefficient above 0.90 can sometimes indicate that several items have highly similar content and therefore provide limited incremental information. In the present study, this possibility is partly mitigated by the IRT findings, which show variation in item discrimination and threshold parameters. Nevertheless, Cronbach's alpha alone cannot establish that the instrument is unidimensional or that the items are nonredundant. The CFA and GRM results are therefore important complements to the reliability coefficient because they provide information about structural validity and item-level functioning that alpha cannot provide.

The GRM results represent an important methodological contribution of the study. All reported discrimination parameters were relatively high, ranging from 1.75 to 2.45. Based on Baker's (2001) classification, these values indicate strong to very strong item discrimination. B1, with an a parameter of 2.45, was the most discriminating item among those reported. This means that the item has a relatively strong capacity to differentiate respondents located at different levels of the underlying trait. In contrast, B30, with an a parameter of 1.75, showed the lowest discrimination among the reported items, although it remained within a high-discrimination range.

The variation in item discrimination demonstrates one of the important advantages of IRT over a purely CTT-based evaluation. Whereas an overall reliability coefficient summarizes the consistency of the entire instrument, the GRM provides information about how individual items contribute to measurement precision. Reise & Moore (2023) emphasize that IRT allows the relationship between latent traits and item responses to be examined at the item level, while Baker (2001) provides the foundational interpretation of discrimination parameters. In the present study, this information makes it possible to identify items that are particularly effective in differentiating students across levels of honesty and discipline and to identify items that may require further qualitative examination.

The threshold parameters provide another layer of diagnostic information. The reported thresholds ranged from -2.12 to 0.55 , indicating that the selected items provide substantial information in the lower-to-middle portion of the latent-trait continuum. B1, for example, had a relatively low threshold, indicating that its response categories become informative at comparatively low levels of the latent trait. Conversely, B30 had a higher threshold, meaning that higher levels of the latent trait were required before respondents were increasingly likely to select higher response categories. This pattern illustrates how the GRM can distinguish not only between highly and poorly discriminating items but also between items that function at different locations along the latent continuum.

The item-fit results further strengthen the evidence supporting the GRM. The finding that all items had $S-X^2$ p-values greater than 0.05 indicates that there was no statistically significant evidence that the observed item-response patterns deviated from the expectations of the model. This supports the appropriateness of the GRM for the response data. The finding is theoretically consistent with the broader argument of Reise & Moore (2023) that IRT provides a model-based framework for evaluating the functioning of individual items, rather than relying exclusively on aggregate test statistics. It is also aligned with the contemporary movement toward more detailed item-level psychometric evaluation in educational and psychological measurement (Belzak, 2023).

The test information function provides an especially important practical interpretation of the instrument's quality. Figure 1 indicates that the instrument provides its greatest measurement information approximately between $\theta = -2.5$ and $+1.0$, with correspondingly lower SEM values in the same region. This means that the instrument is most precise when estimating students whose honesty and discipline are located within this range of the latent continuum. The finding is particularly relevant for educational applications because the instrument may be useful for identifying students whose character-related dispositions are still developing and who may benefit from targeted support or character-building interventions.

This pattern also clarifies the intended function of the instrument. The instrument appears to be more suitable for diagnostic or screening purposes than for high-stakes ranking of students at the upper end of the character continuum. The relatively limited information at high θ levels suggests that the current item set does not contain sufficient high-threshold items to sharply distinguish students who already demonstrate very high levels of honesty and discipline from those who demonstrate moderately high levels. This interpretation is important because measurement precision is not necessarily uniform across the latent-trait continuum. Reise & Moore (2023) emphasize that IRT makes such differences in measurement precision visible through

information functions, an advantage that is difficult to obtain from a single global reliability coefficient.

The lower precision at the upper end of the continuum also provides a clear direction for future instrument refinement. Additional items with higher threshold parameters could be developed to provide greater information for students with very high levels of honesty and discipline (Zhang et al., 2025). Such items would allow the instrument to perform more effectively if its intended purpose were expanded from early identification and intervention toward broader diagnostic profiling or the assessment of advanced character development (Janus & Offord, 2007). In other words, the present findings should not be interpreted simply as evidence of a weakness in the instrument; rather, they identify precisely where additional items would add measurement value.

Taken together, the findings demonstrate that the integration of expert content validation, CFA, internal-consistency analysis, and GRM-based item analysis provides a more comprehensive evaluation of an affective assessment instrument than relying on a single validity or reliability coefficient. The approach responds directly to the limitations of conventional character assessment, in which teacher observations may be subjective and instruments may provide limited information about individual item functioning. The use of IRT makes it possible to identify how strongly individual items discriminate among students and where along the latent continuum the instrument provides the greatest information, thereby offering a more detailed basis for instrument refinement and interpretation (Baker, 2001; Reise & Moore, 2023).

From an educational perspective, the developed instrument has the potential to support evidence-based character assessment in junior secondary schools. Previous studies have emphasized the importance of cultivating discipline and responsibility through educational practices and teacher modeling (Rosita et al., 2022; Saputra et al., 2024; Uge et al., 2022). However, effective character-development programs also require credible assessment mechanisms for determining students' current conditions and evaluating whether interventions produce meaningful changes. The present study contributes to this need by providing an instrument whose quality has been examined not only at the content level but also through empirical structural and item-level analyses.

Overall, the findings suggest that the developed instrument is a promising measurement tool for assessing honesty and discipline among junior secondary school students, particularly for diagnostic purposes. Its strong content validity, satisfactory two-factor structure, high internal consistency, strong item discrimination, acceptable GRM fit, and concentration of measurement information in the lower-to-middle latent-trait range collectively support its psychometric quality. At the same time, the limited information at higher levels of the latent traits indicates that further development is warranted, particularly through the addition of items with higher threshold parameters. Future research should also examine the instrument with larger and more diverse samples, test its measurement invariance across relevant student groups, and cross-validate the GRM structure in independent samples. Such steps would further strengthen the generalizability and practical utility of the instrument for character assessment in Indonesian junior secondary education.

Conclusions

This study developed and validated an IRT-based affective assessment instrument for measuring honesty and discipline among junior secondary school

students. The findings indicate that the developed instrument has strong content and construct validity and demonstrates high internal consistency. The CFA results support the proposed two-dimensional structure, confirming that honesty and discipline can be empirically distinguished as related but separate character constructs. The GRM analysis further indicates that the items demonstrate strong discrimination and satisfactory model fit, while the test information function shows that the instrument provides relatively high measurement precision across the lower-to-middle levels of the latent traits. Overall, the findings suggest that the instrument is psychometrically appropriate for assessing students' honesty and discipline, particularly for diagnostic and early-screening purposes.

The study contributes to educational measurement and character education by demonstrating the usefulness of the graded response model for developing and evaluating Likert-type instruments designed to measure affective characteristics. Unlike conventional approaches that primarily provide overall reliability estimates, the IRT framework enables more detailed evaluation of individual item functioning and measurement precision across different levels of the latent traits. The instrument therefore provides a more systematic and evidence-based alternative to character assessment based predominantly on subjective observation. The findings also support the interpretation of honesty and discipline as distinct dimensions, allowing educators to obtain more meaningful information about students' specific character-related strengths and areas requiring improvement. In practice, the instrument may support teachers and schools in identifying students who require further character development and in informing more data-driven character education programs.

Nevertheless, this study has several limitations. The participants were drawn from junior secondary schools in a particular geographical context, which may limit the generalizability of the findings to students from other regions and educational settings. In addition, the current instrument provides less information at the upper end of the latent-trait continuum, indicating that its ability to differentiate among students with very high levels of honesty and discipline remains limited. Future research should therefore replicate and cross-validate the instrument using larger and more diverse samples from different regions and school contexts. Further item development should also focus on adding items with higher threshold characteristics to improve measurement precision among students with high levels of the assessed traits. Future studies may additionally examine measurement invariance across relevant student groups and investigate the instrument's predictive and criterion-related validity to further establish its fairness, generalizability, and practical usefulness in Indonesian character education.

References

- Ahmad, N. A., Mayouf, A. A., Elias, N. F., & Mohamed, H. (2024). Learning management system instrument development based on Aiken's V technique. *International Journal of Evaluation and Research in Education (IJERE)*, 13(5), 3211. <https://doi.org/10.11591/ijere.v13i5.28925>
- Ashwin, J., Chhabra, A., & Rao, V. (2026). Using Large Language Models for Qualitative Analysis can Introduce Serious Bias. *Sociological Methods & Research*, 55(3), 795–839. <https://doi.org/10.1177/00491241251338246>
- Baker, F. B. (2001). *The Basics of Item Response Theory* (2nd ed.). ERIC Publications.
- Belzak, W. C. M. (2023). The Multidimensionality of Measurement Bias in High-Stakes Testing: Using Machine Learning to Evaluate Complex Sources of Differential Item

- Functioning. *Educational Measurement: Issues and Practice*, 42(1), 24–33. <https://doi.org/10.1111/emip.12486>
- Bowers, D. S., & Sabin, M. (2024). Demonstrating the use of a professional skills framework to support the assessment of dispositions in IT education. *Education and Information Technologies*, 29(6), 7595–7632. <https://doi.org/10.1007/s10639-023-11933-z>
- Carolis, L. De, Kang, I., & Jeon, M. (2026). A Latent Space Graded Response Model for Likert-Scale Psychological Assessments. *Multivariate Behavioral Research*, 61(2), 357–382. <https://doi.org/10.1080/00273171.2025.2605678>
- Carrera, P. G., Ruiz, R. O., Camacho, A., Garandeanu, C. F., Salmivalli, C., & Romera, E. M. (2026). Trajectories of Bystander Behaviors in Bullying during Secondary Education: the Role of Moral Disengagement and Conformity To Peer Pressure. *Journal of Youth and Adolescence*, 55(1), 168–183. <https://doi.org/10.1007/s10964-025-02276-8>
- Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>
- Contreras, C. R., Alarcón, J. C., Soto, S. P., Aguirre, J. B., Liabeuf, G., & Suárez, L. B. (2025). Design and Validation of the Index of Adherence to the Dietary Guidelines for Chile 2022 (GABAS-Index 17). *Nutrients*, 17(22), 3621. <https://doi.org/10.3390/nu17223621>
- Cooper, B., Cohen, T. R., Huppert, E., Levine, E. E., & Fleeson, W. (2023). Honest Behavior: Truth-Seeking, Belief-Speaking, and Fostering Understanding of the Truth in Others. *Academy of Management Annals*, 17(2), 655–683. <https://doi.org/10.5465/annals.2021.0209>
- Gempp, R., & Carrasco, M. G. (2025). Psychometric evaluation of the Brief Multidimensional Students' Life Satisfaction Scale (BMSLSS) in Chilean early adolescents: an Item Response Theory analysis. *Frontiers in Psychology*, 16. <https://doi.org/10.3389/fpsyg.2025.1638017>
- Gustiawan, W., Putri, C. E., Santika, E., Wirda, F., & Suryani, Y. (2025). Beyond Rules: A Comparative View on Work Discipline Practices in Public and Private Workplaces. *Inovbiz: Jurnal Inovasi Bisnis*, 13(1), 01–06. <https://doi.org/10.35314/inovbiz.v13.i1.609>
- Gyamfi, A., & Acquaye, R. (2023). Parameters and Models of Item Response Theory (IRT): A Review of Literature. *Acta Educationis Generalis*, 13(3), 68–78. <https://doi.org/10.2478/atd-2023-0022>
- Hasriani, H., Hamim, U., & Mahmud, R. (2025). Strengthening Scouting Character Values through a Moral Integrated Approach for Elementary School Students. *Indonesian Journal of Innovation Studies*, 26(3). <https://doi.org/10.21070/ijins.v26i3.1463>
- Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, 39(1), 1–22. <https://doi.org/10.1037/cjbs2007001>
- Jeynes, W. H. (2019). A Meta-Analysis on the Relationship Between Character Education and Student Achievement and Behavioral Outcomes. *Education and Urban Society*, 51(1), 33–71. <https://doi.org/10.1177/0013124517747681>
- Jin, K.-Y., Wu, Y.-J., & Chen, H.-F. (2022). A New Multiprocess IRT Model With Ideal Points for Likert-Type Items. *Journal of Educational and Behavioral Statistics*, 47(3), 297–321. <https://doi.org/10.3102/10769986211057160>

- Mansour, W. El, Yoon, M., & Yoon, S. W. (2026). Advancing HRD measurement: a comprehensive guide to IRT and GRM in an international context. *Human Resource Development International*, 29(1), 7–32. <https://doi.org/10.1080/13678868.2025.2558568>
- Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*. Mitra Cendikia Press.
- Monita, D., & Hasan, M. F. (2023). The Dynamics of Affective Assessment in MI Schools: Challenges and Solutions. *Ideguru: Jurnal Karya Ilmiah Guru*, 9(1), 112–119. <https://doi.org/10.51169/ideguru.v9i1.769>
- Paap, K. R., Jefferson, R. T. A., Balakrishnan, N., & Majoubi, J. B. (2023). The many foibles of Likert scales challenge claims that self-report measures of self-control are better than performance-based measures. *Behavior Research Methods*, 56(2), 908–933. <https://doi.org/10.3758/s13428-023-02089-2>
- Patri, R., Madhavan, V., K., V. P., & Manayath, D. (2025). Truthfulness in business managers: views on perception of integrity, ethical challenges, and strategic responses. *Asian Journal of Business Ethics*, 14(1), 121–151. <https://doi.org/10.1007/s13520-025-00231-1>
- Ramadhani, T., Widiyanta, D., Sumayana, Y., Santoso, R. Y., Agustin, P. D., & Al-Amin, A.-A. (2024). The Role Of Character Education In Forming Ethical And Responsible Students. *IJGIE (International Journal of Graduate of Islamic Education)*, 5(2), 110–124. <https://doi.org/10.37567/ijgie.v5i2.3064>
- Reise, S. P., & Moore, T. M. (2023). Item response theory. In *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics (Vol. 1) (2nd ed.)*. (pp. 809–835). American Psychological Association. <https://doi.org/10.1037/0000318-037>
- Rosita, D., Sutisnawati, A., & Uswatun, D. A. (2022). Pendidikan Karakter Nilai Disiplin Dan Tanggung Jawab Dalam Pembelajaran Tematik Di Sekolah Dasar. *Jurnal Cakrawala Pendas*, 8(2), 449–456. <https://doi.org/10.31949/jcp.v8i2.2274>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 100. <https://psycnet.apa.org/record/1972-04809-001>
- Saputra, D. T., Wulandari, M. D., & Darsinah, D. (2024). Penanaman Karakter Disiplin Peserta Didik Melalui Keteladanan Guru di Sekolah Dasar. *Jurnal Basicedu*, 8(1), 99–109. <https://doi.org/10.31004/basicedu.v8i1.6838>
- Sathyanarayana, S., & Mohanasundaram, T. (2024). Fit Indices in Structural Equation Modeling and Confirmatory Factor Analysis: Reporting Guidelines. *Asian Journal of Economics, Business and Accounting*, 24(7), 561–577. <https://doi.org/10.9734/ajeba/2024/v24i71430>
- Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: best practice guidelines for their development and validation. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1494261>
- Tan, P. S. I. (2025). The Influence of Character Education and Family in Preparing Future Leaders Through Christian Education and P5 Practices. *Jurnal Ilmiah Pendidikan Dan Pembelajaran*, 9(2), 296–309. <https://doi.org/10.23887/jipp.v9i2.94111>
- Uge, S., Arisanti, W. O. L., & Hikmawati, H. (2022). Upaya Guru Dalam Menanamkan Karakter Disiplin Siswa Sekolah Dasar. *ELSE (Elementary School Education Journal): Jurnal Pendidikan Dan Pembelajaran Sekolah Dasar*, 6(2), 460. <https://doi.org/10.30651/else.v6i2.13671>
- Zhang, L., Amos, C., & Pentina, I. (2025). Interplay of rationality and morality in using

- ChatGPT for academic misconduct. *Behaviour & Information Technology*, 44(3), 491–507. <https://doi.org/10.1080/0144929X.2024.2325023>
- Zitzmann, S., Bardach, L., Horstmann, K. T., Ziegler, M., & Hecht, M. (2024). Quantifying Individual Personality Change More Accurately by Regression-Based Change Scores. *Structural Equation Modeling: A Multidisciplinary Journal*, 31(5), 909–922. <https://doi.org/10.1080/10705511.2023.2274800>