Indonesian Journal of Educational Development (IJED)
Volume 7, Issue 1, 2026, pp. 1-12
ISSN: 2722-1059 (Online); ISSN: 2722-3671 (Print)
DOI: https://doi.org/10.59672/ijed.v7i1.6114

Gender fairness in measuring student agency: A Rasch DIF analysis

Novi Sylvia*)[1], Ahman[2], Deni Hadiana[3]
[1]Universitas Pendidikan Indonesia, Bandung, Indonesia; novi.sylvia@upi.edu
[2]Universitas Pendidikan Indonesia, Bandung, Indonesia; ahman@upi.edu
[3]Badan Riset dan Inovasi Nasional, Jakarta, Indonesia; deni.hadiana@brin.go.id
*)Corresponding author: Novi Sylvia; E-mail addresses: novi.sylvia@upi.edu

Article Info

Article history:
Received February 05, 2026
Revised February 20, 2026
Accepted March 06, 2026
Available online May 20, 2026

Keywords: Differential item functioning, Gender fairness, Islamic secondary school, Rasch model, Student agency

Abstract. As student agency increasingly shapes educational policy and reform, ensuring that its assessment practices are fair across genders has become a pressing equity concern. This study examines the gender-based measurement fairness of a multidimensional student agency instrument using Rasch-based differential item functioning (DIF) analysis in an Islamic secondary school. The study involved 601 students in Grades 10–12 (301 male; 300 female) selected through a school-based total accessible sampling approach, comprising 96 male and 101 female students in Grade 10, 138 male and 124 female students in Grade 11, and 67 male and 75 female students in Grade 12. Data were collected using a 60-item student agency questionnaire measured on a four-point Likert scale covering eight regulatory, motivational, and future-oriented dimensions. Results show that the instrument operates largely equivalently for male and female students. Although several items exhibited statistically significant DIF, substantively meaningful DIF was limited, localised, and bidirectional, and did not accumulate at the dimensional level. Core regulatory dimensions demonstrated strong invariance across gender. These findings indicate that observed gender differences in agency scores are unlikely to reflect measurement bias but rather reflect contextual variations in the expression of agency within this educational setting.

Introduction

Student agency has emerged as a cornerstone of contemporary educational discourse, reflecting a broader shift from transmission-oriented models of schooling toward learner-centred, self-directed, and future-oriented education (Cook-Sather, 2020). Across global policy frameworks and academic research, student agency is increasingly positioned as a foundational capacity enabling learners to navigate complex learning environments, regulate their own learning processes, and adapt to rapidly changing social and economic contexts (Lee, 2025). From this perspective, agency is not merely an individual trait but a dynamic interaction among learners, instructional practices, and sociocultural environments that shape how students engage with learning opportunities and constraints (Adhikari, 2024a).

Theoretically, student agency is rooted in social cognitive theory, which emphasises intentionality, self-regulation, and forethought as central features of human functioning (Bandura, 2001; Bandura, 2006). In educational psychology, agency is closely associated with constructs such as self-efficacy, goal orientation, persistence, and metacognitive regulation, all of which enable learners to exercise control over their learning trajectories (Zeiser et al., 2018). More recent perspectives extend this

individual-focused view by situating agency within broader institutional and cultural contexts, highlighting how pedagogical practices, assessment regimes, and sociocultural norms afford or constrain opportunities for agentic action (Marín et al., 2025). Consequently, student agency is increasingly understood as both a personal capacity and a contextually enacted phenomenon.

As interest in student agency continues to expand, so too does the need for valid, reliable, and fair measurement instruments. Educational decisions ranging from classroom-level interventions to system-wide policy reforms rely on data derived from psychometric measures of learner dispositions and competencies (Ackerman et al., 2024; Deakin Crick et al., 2015). When such instruments function differently across groups, observed score differences may reflect measurement bias rather than genuine disparities in underlying capacities. Measurement invariance, therefore, represents a central concern in contemporary educational assessment, particularly for constructs increasingly invoked in equity-oriented policies (Wu et al., 2017).

One of the most rigorous approaches to evaluating measurement invariance is differential item functioning (DIF) analysis within the Rasch measurement framework. DIF analysis examines whether individuals from different groups, but with equivalent levels of the underlying latent trait, have different probabilities of endorsing particular items (Boone & Staver, 2020). When DIF is present, group comparisons based on raw or scaled scores may be misleading, potentially reinforcing inequitable interpretations and policy decisions (Russell & Kaplan, 2021). Rasch-based DIF analysis is particularly appropriate for complex, multidimensional constructs such as student agency because it provides invariant item estimates and emphasises construct coherence (Boone et al., 2014).

Despite the centrality of student agency in educational theory and practice, empirical studies examining DIF in student agency instruments remain limited. Existing research has predominantly focused on construct validation, factor structure, and reliability (Adhikari, 2024a; Brandt, 2024; Cavazzoni et al., 2022; Jo et al., 2022), with relatively few studies subjecting agency measures to rigorous invariance testing across demographic groups. Moreover, much of the DIF and measurement invariance literature has concentrated on cognitive outcomes rather than motivational or agentic constructs (Kurnaz & Yildiz, 2023), leaving a critical gap in the psychometric evaluation of agency measures. This gap is further compounded by the dominance of Western samples in psychometric research, raising questions about the generalizability and fairness of agency instruments across diverse sociocultural and educational systems (Ye, 2024).

Gender is among the most frequently examined grouping variables in educational research, particularly in studies of motivation, self-regulation, and academic engagement. Numerous studies report gender differences in agency-related constructs, often attributing these patterns to socialisation processes, instructional practices, or cultural expectations (Firdoos et al., 2023; Pumbay, 2018; Storms, 2019). Prior research suggests that gender differences in motivational and self-regulatory constructs are often smaller and more context-dependent than commonly assumed, aligning with the gender similarities hypothesis Gegenfurtner (2020) and meta-analytic evidence on self-efficacy and achievement-related beliefs (Nabunya et al., 2021; Yu & Deng, 2022).

However, contemporary social psychological research shows that gender differences in agency are complex and multidimensional rather than uniform. Agency is commonly contrasted with communality, where agency refers to attributes such as assertiveness, independence, dominance, and achievement orientation, whereas communality encompasses warmth, cooperation, and relational orientation. Across many countries, men tend to rate themselves slightly higher in agentic traits than women, yet these differences are generally small in magnitude. Large-scale cross-national analyses report average gender differences around $d \approx 0.20$, and statistically significant gaps appear

only in a subset of countries examined (Kosakowska-Berezecka et al., 2023). Importantly, perceived gender differences are often larger at the level of social stereotypes than in actual self-reports, with public perceptions consistently portraying men as more agentic while viewing women as more communal, even when both genders are perceived as similarly competent and capable (Eagly et al., 2020).

Further nuance emerges when agency is decomposed into multiple components rather than treated as a single dimension. Multidimensional research distinguishes among instrumental competence, leadership orientation, assertiveness, and independence. Findings consistently show that women are typically perceived as less dominant or assertive in stereotypical portrayals, yet comparable to men in competence and independence (Folberg et al., 2022; Hentschel et al., 2019). Self-assessment patterns similarly indicate that women do not rate themselves lower in competence or autonomy, though they may report slightly lower levels of leadership or assertiveness (Hentschel et al., 2019). Moreover, large workplace studies reveal minimal gender differences in agentic behaviours once occupational context is controlled, with the most stable differences instead appearing in communal traits, where women consistently score higher (Gartzia, 2022).

These patterns highlight the importance of contextual and normative influences in shaping how agency is expressed and perceived. Social role theory posits that gender stereotypes emerge from historically patterned distributions of men and women across social and occupational roles (Suhardita et al., 2024). Because men have traditionally occupied more competitive and high-status positions, agentic characteristics become culturally associated with masculinity, while communal traits become linked to femininity (Ellemers, 2018; Gustafsson Sendén et al., 2019; Sczesny et al., 2018). Such stereotypes, in turn, influence expectations and evaluations: candidates for prestigious or competitive roles are often perceived as more suitable when they display strong agentic traits (Dutz et al., 2022). However, women who express highly dominant forms of agency may face social penalties, a phenomenon often described as the "agentic penalty," where competence is acknowledged, but social likability decreases (Gustafsson Sendén et al., 2019; Ma et al., 2022; Rosette et al., 2016).

Importantly, these normative pressures do not necessarily reflect inherent differences in agentic capacities but rather socially conditioned expectations regarding appropriate gender behaviour. Historical analyses of literary and media portrayals, for instance, show that female characters have long been depicted as more passive relative to male characters, though such gaps have narrowed over time (Stuhler, 2024). Cross-national research further indicates that in societies with greater gender equality and lower power-distance norms, gender differences in agency tend to shrink, often because men report lower dominance claims rather than because women dramatically increase agentic self-perceptions (Kosakowska-Berezecka et al., 2023). These findings collectively suggest that gender differences in agency are highly sensitive to cultural, institutional, and normative contexts.

Despite these nuanced insights, gender differences in agency-related constructs are often interpreted without sufficient attention to measurement fairness. Observed differences are frequently attributed to socialisation processes, instructional practices, or cultural norms, without first establishing whether the instruments themselves function equivalently for male and female students. This issue becomes especially salient in educational contexts subject to normative assumptions or social stigma, such as religious schooling. Islamic secondary schools, in particular, are frequently portrayed in public discourse as environments that may reinforce traditional gender roles or constrain female agency. While such claims are often grounded in ideological or cultural critiques, they are rarely examined through the lens of psychometric evidence. As a result, debates

surrounding gender and Islamic education tend to rely more on assumptions than on systematic empirical validation.

Importantly, Islamic schooling is not monolithic. In Indonesia, a Muslim-majority nation, Islamic secondary schools operate within diverse pedagogical, cultural, and policy frameworks. Many Islamic schools integrate national curricula with religious and moral education, explicitly emphasising values such as responsibility, self-discipline, moral reasoning, and self-regulation (Guna et al., 2024; Winda & Surawan, 2025). These values align closely with contemporary conceptualisations of student agency, suggesting that religious schooling may cultivate, rather than constrain, agentic capacities (Shim, 2021). Yet, the extent to which student agency can be measured fairly across gender within these educational settings remains empirically underexplored.

Responding to these gaps, the present study addresses the following research question: Does the student agency instrument demonstrate measurement invariance across male and female students when evaluated using Rasch-based differential item functioning (DIF) analysis? Based on contemporary research indicating that gender differences in agency-related competencies are generally small and highly context-dependent, the study is guided by the hypothesis that the instrument will function equivalently across genders. Accordingly, the objective of this study is to evaluate the measurement fairness of the student agency instrument across gender groups and to determine whether observed score differences reflect genuine variation in agency-related dispositions rather than measurement artefacts.

Method

Research Design

This study employed a quantitative, cross-sectional design grounded in Rasch measurement theory Bond & Fox (2015); Boone & Staver (2020); Creswell & Creswell (2022) to examine the gender-based measurement fairness of a multidimensional student agency instrument. Rasch-based differential item functioning (DIF) analysis was selected because it provides invariant item estimates across samples, supports principled evaluation of measurement equivalence, and enables transparent interpretation of group-related item performance through a common measurement scale (Bond & Fox, 2015; Boone et al., 2014). DIF was treated as a primary analytic objective rather than a post hoc diagnostic, consistent with contemporary recommendations emphasising fairness testing as an integral component of instrument validation (Wu et al., 2017).

Participants and Sampling Technique

The study employed a school-based total accessible sampling approach. All available students in Grades 10 and 11 at the selected Islamic secondary school participated in the study, while participation from Grade 12 included students from several classes who were accessible during the data collection period. In total, 601 students were included in the analysis. The sample consisted of 301 male and 300 female students, ensuring near-balanced gender representation. In Grade 10, participants included 96 male and 101 female students. Grade 11 comprised 138 male and 124 female students, while Grade 12 included 67 male and 75 female students. This distribution provided substantial representation across grade levels and gender groups within the school, thereby supporting stable Rasch calibration and DIF analysis across gender.

Data Collection Technique and Research Instrument

Data were collected using a self-administered student agency questionnaire distributed during scheduled school sessions under teacher supervision. All data collection procedures complied with institutional ethical standards. This study obtained ethical approval from the Indonesia University of Education under reference number 12/UN40.K/PT.01.01/2026. Participation was voluntary,

and informed consent was obtained prior to data collection. Students were informed that responses would be anonymised and used solely for research purposes.

The study employed a 60-item student agency questionnaire grounded in established multidimensional models of learner agency and adapted from Zeiser et al. (2018). The instrument comprised eight theoretically distinct dimensions: self-efficacy (8 items), persistence of interest (6 items), perseverance of effort (6 items), locus of control (10 items), mastery orientation (5 items), metacognitive self-regulation (10 items), self-regulated learning (9 items), and future orientation (6 items). All items were rated on a four-point Likert-type scale ranging from 1 (strongly disagree) to 4 (strongly agree), using an ordered polytomous response format appropriate for Rasch rating scale analysis. The absence of a neutral midpoint was intended to encourage directional responses and to reduce central-tendency bias.

Each dimension was analysed separately to preserve construct clarity and satisfy Rasch model assumptions, particularly unidimensionality within each scale (Bond & Fox, 2015; Boone et al., 2014). This dimensional approach aligns with recommendations for assessing complex motivational and regulatory constructs, where aggregating heterogeneous subdimensions may obscure meaningful variance and weaken interpretability. Collectively, the eight dimensions provided a theoretically grounded and psychometrically appropriate representation of student agency for Rasch-based calibration and subsequent DIF analysis.

Preliminary Rasch Analyses

Prior to DIF analysis, a series of Rasch analyses were conducted to evaluate the psychometric adequacy of each dimension. These analyses examined item fit, person and item reliability, separation indices, category functioning, and dimensionality (Boone & Staver, 2020; Linacre, 2023). Item fit was evaluated using infit and outfit mean square statistics, with values within commonly accepted ranges indicating adequate model-data fit and suggesting that items contribute meaningfully to the measurement of the latent construct (Bond & Fox, 2015). Person and item reliability coefficients were inspected to assess measurement precision across respondents and items, respectively. Dimensionality was assessed using principal component analysis (PCA) of Rasch residuals, a standard procedure for evaluating essential unidimensionality within Rasch measurement (Boone et al., 2014; Linacre, 2023). Evidence of essential unidimensionality was established when the proportion of variance explained by the Rasch dimension was substantively adequate and when the eigenvalue of the first residual contrast remained below conventional thresholds. These criteria collectively supported the suitability of each dimension for subsequent DIF analysis.

Differential Item Functioning (DIF) Analysis

Differential item functioning (DIF) was examined using the Rasch measurement model to evaluate measurement invariance across gender. The Rasch framework enables invariant item calibration across groups while controlling for differences in the latent trait, thereby allowing direct examination of item-level bias (Bond & Fox, 2015; Boone & Staver, 2020). Within this framework, DIF analysis assesses whether individuals from different groups but with equivalent latent trait levels demonstrate systematically different probabilities of endorsing specific items (Russell & Kaplan, 2021). Meaningful DIF therefore indicates potential measurement non-invariance rather than true group differences.

Prior to conducting DIF analyses, Rasch measurement evaluations confirmed adequate person and item reliability, acceptable item fit, and essential unidimensionality for each dimension. Establishing these psychometric prerequisites is critical because meaningful DIF interpretation presupposes that each scale functions as a coherent unidimensional measure of the intended construct (Boone et al.,

2014). DIF analyses were conducted at the item level within each dimension, consistent with best practices for analysing multidimensional constructs, as this prevents the conflation of heterogeneous latent traits and allows precise identification of sources of measurement non-invariance (Bond & Fox, 2015). All DIF analyses were performed using Winsteps software (Linacre, 2023). DIF magnitude was quantified using Rasch DIF contrasts expressed in logits, calculated as differences in item difficulty estimates between female and male groups. Positive DIF contrast values indicate that an item is relatively more difficult for female students, whereas negative values indicate greater difficulty for male students. This directional interpretation facilitates substantive understanding of item-level gender effects without presuming advantage or disadvantage at the scale level.

DIF Decision Rules and Interpretation

In accordance with established Rasch measurement guidelines, DIF interpretation prioritised effect size alongside statistical significance (Boone et al., 2014; Russell & Kaplan, 2021). DIF contrasts below 0.43 logits were considered negligible, contrasts between 0.43 and 0.63 logits moderate, and contrasts of 0.64 logits or higher large. These thresholds are widely applied in Rasch-based DIF research and reflect practical, not purely statistical, considerations of measurement non-invariance. The statistical significance of DIF contrasts was evaluated using Welch's t-statistics and associated p-values generated by Winsteps. An alpha level of $p < .05$ was adopted. However, consistent with contemporary recommendations, statistical significance alone was not considered sufficient evidence of meaningful DIF, particularly in large samples where trivial differences may reach significance (Russell & Kaplan, 2021; Wu et al., 2017). An item was flagged as exhibiting substantive DIF only when both statistical significance and effect-size thresholds were satisfied. Items meeting statistical but not effect-size criteria were classified as demonstrating statistical DIF without substantive impact. This combined decision rule mitigates over-identification of DIF and ensures focus on practically meaningful measurement departures.

DIF results were examined both at the item and dimensional levels. Dimensions were considered free from meaningful DIF when substantively flagged items were few, bidirectional, and non-cumulative. This interpretation reflects the view that localised DIF does not necessarily threaten overall measurement fairness, especially for complex motivational constructs such as student agency (Boone & Staver, 2020). DIF decisions were made at both item and dimensional levels. Dimensions were considered measurement invariant when substantively meaningful DIF was absent or when flagged items were limited, bidirectional, and non-cumulative. This approach aligns with contemporary fairness perspectives, emphasising interpretive caution and contextualised evaluation rather than rigid elimination of all DIF instances (Wu et al., 2017).

Analytic Rationale

The analytic strategy reflects a balance between methodological rigour and interpretive caution. Establishing psychometric adequacy prior to DIF analysis, applying effect-size-informed thresholds, and evaluating patterns at both the item and dimensional levels minimise the risk of spurious findings while maintaining sensitivity to meaningful sources of bias. Importantly, the focus on measurement fairness does not assume the absence of gendered experiences. Rather, it distinguishes substantive differences in agency from artefacts arising from the functioning of instruments. This distinction is critical in contexts subject to normative assumptions, such as religious schooling, where claims of bias should be grounded in empirical measurement evidence rather than ideological presuppositions.

Results and Discussion

The primary objective of this study was to determine whether the student agency instrument demonstrates measurement invariance across gender within an Islamic secondary school context. In contemporary educational discourse, student agency increasingly functions not only as a pedagogical aspiration but also as a measurable outcome informing policy and intervention (Ackerman et al., 2024; Lee, 2025; Marín et al., 2025). Because such measurements may influence instructional decisions and resource allocation, ensuring fairness across demographic groups is methodologically and ethically essential. To address this concern, Rasch-based differential item functioning (DIF) analysis was employed to examine whether male and female students with equivalent levels of the latent trait responded differently to specific items.

Within the Rasch framework, DIF occurs when individuals from different groups, despite having the same underlying trait level, show unequal probabilities of endorsing an item (Boone et al., 2014; Boone & Staver, 2020; Russell & Kaplan, 2021). Such discrepancies indicate potential threats to measurement invariance because group differences in observed scores may reflect item bias rather than substantive variation in the construct itself (Wu et al., 2017). Thus, DIF analysis serves as a diagnostic tool to distinguish genuine trait differences from artefacts of instrument functioning.

As summarised in Image 1, several items displayed statistically significant DIF. However, only a limited subset met thresholds for substantively meaningful DIF based on combined statistical significance and effect-size criteria. Importantly, substantively flagged items were distributed across multiple dimensions and demonstrated bidirectional contrasts: some favouring male students and others favouring female students. No dimension showed concentrated or cumulative DIF sufficient to distort scale-level interpretation. This distribution pattern is critical: when DIF effects are localised and non-systematic, they indicate contextual sensitivity rather than structural measurement bias.

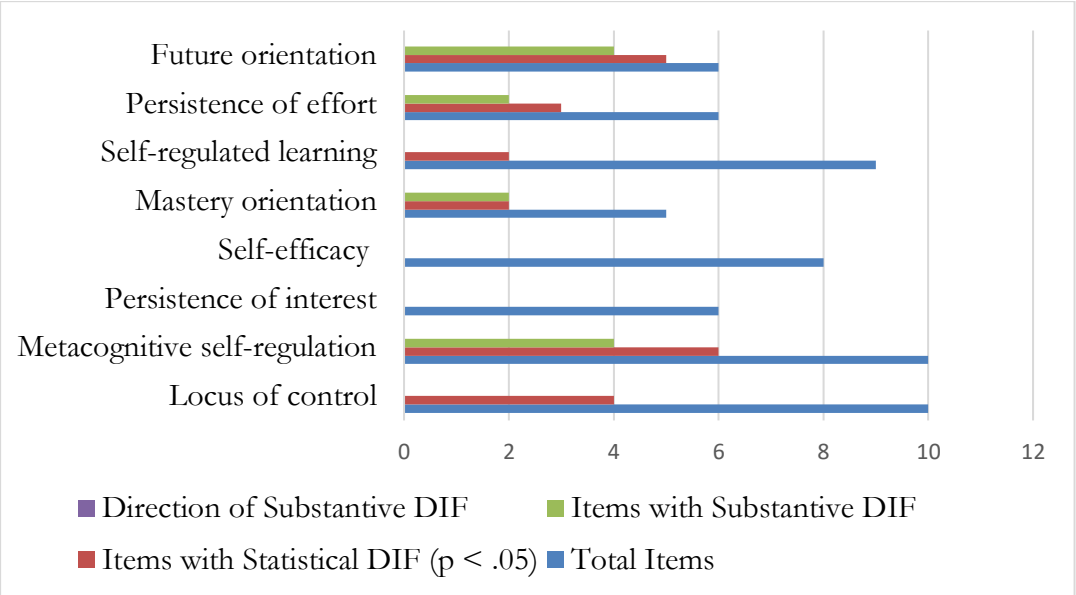

Image 1. Summary of Gender-Based Differential Item Functioning (DIF)

Further support for invariance emerged from Wright maps and person-item targeting analyses generated through Winsteps. The distributions of male and female students substantially overlapped across the agency continuum, indicating comparable representation of trait levels across

gender groups. Moreover, items were appropriately targeted to the range of student ability levels, reducing the likelihood that DIF findings resulted from misalignment between item difficulty and respondent capability. Together, these findings strengthen confidence that the instrument functions equivalently across gender.

Dimensional-Level Findings

Analysis at the dimensional level revealed that locus of control, self-regulated learning, persistence of interest, and self-efficacy demonstrated particularly robust invariance. No substantively meaningful DIF emerged within these dimensions, suggesting that foundational regulatory components of agency are assessed equitably across male and female students. These dimensions represent core elements of agentic functioning: belief in personal influence, strategic regulation of learning, sustained engagement, and perceived competence. Their invariance indicates that the structural backbone of the agency construct remains stable across gender in this context.

This finding aligns with contemporary research suggesting that gender differences in motivational and regulatory constructs are often smaller and more context-dependent than widely assumed (Firdoos et al., 2023; Nabunya et al., 2021; Yu & Deng, 2022; Widana, 2022). When measurement fairness is established, apparent disparities frequently diminish or disappear, underscoring the importance of psychometric validation prior to substantive interpretation. Thus, the present results support the gender similarities hypothesis within the domain of regulatory agency in educational settings.

Localised DIF emerged primarily within future orientation, mastery orientation, metacognitive self-regulation, and perseverance of effort. However, the bidirectional nature of these effects suggests that sensitivity arises from item content rather than construct inequivalence. Items referencing long-term aspirations, goal-setting, or persistence may intersect with socially patterned expectations regarding academic roles, leadership, or responsibility. Such contextual interpretation does not necessarily signal unfairness but reflects the dynamic interplay between agency and sociocultural norms.

Localised DIF as Contextual Sensitivity

Rather than interpreting localised DIF as a flaw, it may be understood as evidence of the contextually enacted nature of agency. Agency is not a static psychological attribute but a capacity exercised within institutional and normative structures (Adhikari, 2024b; Marín et al., 2025). Students may interpret items differently depending on how their environment frames responsibility, ambition, or autonomy. Gendered expectations, particularly regarding leadership assertiveness or future career planning, may influence item-level response patterns without compromising the integrity of the broader construct.

In Rasch measurement practice, a complete absence of DIF across all items is uncommon in multidimensional motivational instruments. The key issue is whether DIF accumulates to distort dimensional meaning. In the present study, it did not. Consequently, localised DIF signals opportunities for refinement, such as clarifying item wording or neutralising contextual phrasing, rather than grounds for rejecting the scale.

Theoretical Implications

Theoretically, these findings reinforce a multidimensional and context-sensitive understanding of student agency. The invariance of core regulatory dimensions supports the conceptualisation of agency as structured around self-regulatory competence rather than dominance-based assertiveness. Moreover, by demonstrating invariance within a religious educational setting, the study extends cross-context validation of agency constructs beyond Western-centric samples (Ye,

2024). This contributes to global measurement scholarship by illustrating that agency frameworks can maintain construct coherence across diverse sociocultural environments.

Importantly, the findings also caution against conflating social stereotypes with psychological measurement. Perceived gender gaps in agency are often amplified in public discourse, yet psychometric evidence suggests that foundational capacities are comparable when measured fairly. Thus, establishing invariance serves not only statistical but also epistemological functions. It prevents normative assumptions from shaping empirical interpretation.

Practical Implications

In practice, the results provide reassurance to educators and policymakers who use agency assessments to inform instructional planning. Because the instrument demonstrates gender fairness, observed differences may be interpreted as meaningful rather than artefact-driven. This strengthens confidence in using agency scores to design interventions targeting motivation, persistence, or strategic learning behaviours.

Furthermore, the findings underscore the importance of routine fairness evaluation in school-based assessment. As psychological constructs increasingly guide educational reform, ensuring equitable measurement across demographic groups becomes a necessary component of responsible policy implementation.

Revisiting Gender Assumptions in Islamic Schooling

By establishing measurement invariance, this study challenges generalised claims that Islamic schooling inherently restricts female agency. The absence of systematic DIF across core dimensions indicates that foundational agentic capacities are assessed equitably in this institutional context. Islamic schooling in Indonesia integrates moral formation with disciplined self-regulation and responsibility (Guna et al., 2024; Winda & Surawan, 2025), aligning with contemporary agency frameworks emphasising purposeful autonomy (Shim, 2021). Thus, assumptions regarding constrained female agency require empirical validation rather than reliance on ideological narratives.

Limitations and Future Directions

Despite robust sample size and rigorous Rasch analysis, several limitations warrant consideration. First, measurement invariance was examined solely across gender within a single school context. While sufficient for stable calibration, the school-based total accessible sampling design limits broader generalizability. Second, only one grouping variable was tested. Other structural factors, such as socioeconomic background, academic track, or regional variation, may shape item interpretation and merit future DIF investigation.

Future research should therefore extend invariance testing to multi-site samples and additional demographic variables. Longitudinal designs may also explore whether invariance remains stable across developmental stages. Such extensions would deepen understanding of how agency operates and is measured within evolving educational environments.

Conclusion

Rasch-based differential item functioning (DIF) analysis confirms that the student agency instrument demonstrates measurement invariance across male and female students in the examined Islamic secondary school, thereby answering the research question and supporting gender-based measurement fairness. Although several items showed statistical DIF, substantively meaningful effects were limited, localised, and bidirectional, and did not accumulate at the dimensional level.

Core regulatory dimensions exhibited strong invariance, indicating that foundational aspects of student agency are assessed equitably across gender. These findings suggest that observed gender differences in agency scores are unlikely to reflect measurement bias but rather reflect contextual variation in the expression of agency. The results provide empirical support for the continued use of the instrument in educational settings and highlight the importance of routine fairness monitoring to ensure valid and equitable interpretation of agency assessments across diverse student populations.

Bibliography

Ackerman, T. A., Bandalos, D. L., Briggs, D. C., Everson, H. T., Ho, A. D., Lottridge, S. M., Madison, M. J., Sinharay, S., Rodriguez, M. C., Russell, M., von Davier, A. A., & Wind, S. A. (2024). Foundational competencies in educational measurement. *Educational Measurement: Issues and Practice*, *43*(3), 7–17. https://doi.org/10.1111/emip.12581

Adhikari, D. P. (2024a). Constructing student agency: The nexus between classroom activities and engagement. *International Journal of Education and Practice*, *12*(3). https://doi.org/10.18488/61.v12i3.3759

Adhikari, D. P. (2024b). Constructing student agency: The nexus between classroom activities and engagement. *International Journal of Education and Practice*, *12*(3), 819–830. https://doi.org/10.18488/61.v12i3.3759

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*(1), 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, *1*(2), 164–180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.)*. Routledge.

Boone, W. J., & Staver, J. R. (2020). Advances in Rasch analyses in the human sciences. In *Advances in Rasch Analyses in the Human Sciences*. https://doi.org/10.1007/978-3-030-43420-5

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

Brandt, W. C. (2024). Measuring student success skills: A review of the literature on student agency. Competencies of the future. In *National Center for the Improvement of Educational Assessment*.

Cavazzoni, F., Fiorini, A., & Veronese, G. (2022). How do we assess how agentic we are? A literature review of existing instruments to evaluate and measure individuals' agency. *Social Indicators Research*, *159*(3), 1125–1153. https://doi.org/10.1007/s11205-021-02791-8

Cook-Sather, A. (2020). Student voice across contexts: Fostering student agency in today's schools. *Theory Into Practice*, *59*(2), 182–191. https://doi.org/10.1080/00405841.2019.1705091

Creswell, J. W., & Creswell, J. D. (2022). *Research Design: qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Deakin Crick, R., Huang, S., Ahmed Shafi, A., & Goldspink, C. (2015). Developing resilient agency in learning: The internal structure of learning power. *British Journal of Educational Studies*, *63*(2), 121–160. https://doi.org/10.1080/00071005.2015.1006574

Dutz, R., Hubner, S., & Peus, C. (2022). When agency "fits" regardless of gender: Perceptions of applicant fit when job and organization signal male stereotypes. *Personnel Psychology*, *75*(2), 441–483. https://doi.org/10.1111/peps.12470

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*(3), 301–315. https://doi.org/10.1037/amp0000494

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, *69*(1), 275–298. https://doi.org/10.1146/annurev-psych-122216-011719

Firdoos, A., Naz, F. L., & Masud, Z. (2023). Impact of cultural norms and social expectations for shaping gender disparities in educational attainment in Pakistan. *Qlantic Journal of Social Sciences and Humanities*, *4*(3), 166–172. https://doi.org/10.55737/qjssh.311246563

Folberg, A. M., Zhu, M., He, Y., & Ryan, C. S. (2022). The primacy of nurturance and dominance/assertiveness: Unidimensional measures of the big two mask gender differences in subdimensions. *International Review of Social Psychology*, *35*(1). https://doi.org/10.5334/irsp.690

Gartzia, L. (2022). Self and other-reported workplace traits: A communal gap of men across occupations. *Journal of Applied Social Psychology*, *52*(8), 568–587. https://doi.org/10.1111/jasp.12848

Gegenfurtner, A. (2020). Testing the gender similarities hypothesis: differences in subjective task value and motivation to transfer training. *Human Resource Development International*, *23*(3), 309–320. https://doi.org/10.1080/13678868.2018.1449547

Guna, B. W. K., Yuwantiningrum, S. E., Firmansyah, S, Muh. D. A., & Aslan, A. (2024). Building morality and ethics through Islamic religious education in schools. *IJGIE (International Journal of Graduate of Islamic Education)*, *5*(1), 14–24. https://doi.org/10.37567/ijgie.v5i1.2685

Gustafsson Sendén, M., Klysing, A., Lindqvist, A., & Renström, E. A. (2019). The (Not so) changing man: dynamic gender stereotypes in sweden. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00037

Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00011

Jo, Y., Park, S., & Jung, W. (2022). Development of tools to measure student agency for middle and high school students. *Korean Association For Learner-Centered Curriculum And Instruction*, *22*(11), 189–211. https://doi.org/10.22251/jlcci.2022.22.11.189

Kosakowska-Berezecka, N., Bosson, J. K., Jurek, P., Besta, T., Olech, M., Vandello, J. A., Bender, M., Dandy, J., Hoorens, V., Jasinskaja-Lahti, I., Mankowski, E., Venäläinen, S., Abuhamdeh, S., Agyemang, C. B., Akbaş, G., Albayrak-Aydemir, N., Ammirati, S., Anderson, J., Anjum, G., … Żadkowska, M. (2023). Gendered self-views across 62 countries: A test of competing models. *Social Psychological and Personality Science*, *14*(7), 808–824. https://doi.org/10.1177/19485506221129687

Kurnaz, F. B., & Yildiz, H. (2023). Investigating the sources of differential item functioning: A sample critical thinking motivation scale. *International Journal of Assessment Tools in Education*, *10*(3), 434–453. https://doi.org/10.21449/ijate.1279152

Lee, S. (2025). Making sense of 'student agency': The subjectivity of the learner in globalized curriculum reform and the case of South Korea. *Journal of Philosophy of Education*, *59*(3–4), 510–526. https://doi.org/10.1093/jopedu/qhaf015

Linacre, J. M. (2023). A User's Guide to Winsteps? Ministep: Rasch-model computer programs. In *Program Manual 5.5.1*.

Ma, A., Rosette, A. S., & Koval, C. Z. (2022). Reconciling female agentic advantage and disadvantage with the CADDIS measure of agency. *Journal of Applied Psychology*, *107*(12), 2115–2148. https://doi.org/10.1037/apl0000550

Marín, V. I., Tur, G., Castañeda, L., Peguera-Carré, M. C., Orellana, M. L., Villagrá, S., & Carrera, X. (2025). Agencia y aprendizaje en la Educación Superior: Una revisión sistemática. *Universitas Tarraconensis Revista de Ciències de l Educació*, (1), e4035. https://doi.org/10.17345/ute.2025.4035

Nabunya, P., Curley, J., & Ssewamala, F. M. (2021). Gender norms, beliefs and academic achievement of orphaned adolescent boys and girls in uganda. *The Journal of Genetic Psychology*, *182*(2), 89–101. https://doi.org/10.1080/00221325.2021.1873727

Pumbay, M. (2018). *Role of gender differences, agency and aspirations in women's medical career decisions: evidence from Pakistan*. International Institute of Social Studies.

Rosette, A. S., Koval, C. Z., Ma, A., & Livingston, R. (2016). Race matters for women leaders: Intersectional effects on agentic deficiencies and penalties. *The Leadership Quarterly*, *27*(3), 429–445. https://doi.org/10.1016/j.leaqua.2016.01.008

Russell, M., & Kaplan, L. (2021). An intersectional approach to differential item functioning: reflecting configurations of inequality. *Practical Assessment, Research and Evaluation*, *26*.

Sczesny, S., Nater, C., & Eagly, A. H. (2018). Agency and communion. In *Agency and Communion in Social Psychology* (pp. 103–116). Routledge. https://doi.org/10.4324/9780203703663-9

Shim, J.-M. (2021). Religiosity and individual agency: Denominational affiliation, religious action, and Sense of Control (SOC) in Life. *Religions*, *12*(2), 117. https://doi.org/10.3390/rel12020117

Storms, C. (2019). Gender Differences: A result of differences in the brain or socialization? *Locus: The Seton Hall Journal of Undergraduate Research*, *2*(1). https://doi.org/10.70531/2573-2749.1018

Stuhler, O. (2024). The gender agency gap in fiction writing (1850 to 2010). *Proceedings of the National Academy of Sciences*, *121*(29). https://doi.org/10.1073/pnas.2319514121

Suhardita, K., Widana, I. W., Degeng, I. N. S., Muslihati, M., & Indreswari, H. (2024). Sharing behavior in the context of altruism is a form of strategy for building empathy and solidarity. *Indonesian Journal of Educational Development (IJED)*, *5*(3), 316-324. https://doi.org/10.59672/ijed.v5i3.4145

Widana, I. W. (2022). Meta-analysis: The relationship between self-regulated learning and mathematical critical reasoning. *Education.Innovation.Diversity, 1*(4), 64-75. https://doi.org/10.17770/eid2022.1.6739

Winda, W., & Surawan, S. (2025). Transformasi pendidikan Islam dalam mendorong self-regulated learning siswa madrasah: Studi empiris di MTs Mumtaz Palangka Raya (Transformation of Islamic education in encouraging self-regulated learning of madrasah students: An empirical study at MTS Mumtaz Palangka Raya). *ARZUSIN*, *5*(3), 1560–1571. https://doi.org/10.58578/arzusin.v5i3.6052

Wu, A. D., Liu, Y., Stone, J. E., Zou, D., & Zumbo, B. D. (2017). Is difference in measurement outcome between groups differential responding, bias or disparity? A methodology for detecting bias and impact from an attributional stance. *Frontiers in Education*, *2*. https://doi.org/10.3389/feduc.2017.00039

Ye, S. (2024). Fundamental attribution error in the classroom: Why and how bias hurts? *Lecture Notes in Education Psychology and Public Media*, *61*(1), 27–34. https://doi.org/10.54254/2753-7048/61/20240427

Yu, Z., & Deng, X. (2022). A meta-analysis of gender differences in e-learners' self-efficacy, satisfaction, motivation, attitude, and performance across the world. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.897327

Zeiser, K., Scholz, C., & Cirks, V. (2018). Maximizing student agency: Implementing and measuring student-centered learning practices. *American Institutes for Research*.