Gadjah Mada Journal of Psychology (GamaJoP)
Journal Homepage: https://jurnal.ugm.ac.id/gamajop
Email : gamajop.psikologi@ugm.ac.id
ISSN 2407 - 7798 (Online)

Volume **12**
Number 2, 2026
Page: 143–151
DOI:10.22146/gamajop.107088

Received 25 May 2025
Revised 24 August 2026
Accepted 24 August 2026
Published 29 October 2026

Keywords:
CFA; DIF; Mediated MIMIC model; self-efficacy

***Author for correspondence:** Email: rahmatsbintang@universitasbosowa.ac.id

The Indonesian Version of the Generalized Self-Efficacy Scale (GSE): Differential Item Functioning Analysis Using a Mediated MIMIC Model

Rahmat S. Bintang,*[,1] Burhan,[2] and Anhar Dana Putra[3]

[1]Faculty of Psychology, Universitas Bosowa, Makassar, Indonesia
[2]Faculty of Education and Literature, Universitas Bosowa, Makassar, Indonesia
[3]Faculty of Business and Economics, Monash University, Melbourne, Australia

Abstract

Self-efficacy refers to an individual's optimistic belief in their own capabilities—the belief that they can perform difficult tasks and overcome challenges effectively. One commonly used instrument for measuring self-efficacy is the Generalized Self-Efficacy Scale (GSE). This study aimed to examine the factor structure and differential item functioning (DIF) of the Indonesian version of the GSE. Confirmatory factor analysis (CFA) was used to test the factor structure, whereas the Mediated MIMIC model was applied to assess DIF. The participants were 534 undergraduate students (102 males and 432 females) aged 17–25 years (M = 20.06, SD = 1.34). A non-probability sampling method with a convenience sampling technique was used. The findings indicated that CFA supported a one-factor (unidimensional) model with a good fit to the data. At the item-parameter level, all 10 items of the Indonesian version of the GSE showed statistically significant factor loadings. However, gender-related DIF was detected in Item 3. These findings suggest that although the Indonesian version of the GSE demonstrates a unidimensional measurement structure, researchers and practitioners should interpret gender-based comparisons cautiously because Item 3 may function differently across male and female respondents.

Introduction

Self-efficacy was first introduced by Albert Bandura (1977) within social cognitive theory, which conceptualizes the interaction among individuals, behavior, and the environment. Bandura referred to self-efficacy as the belief that individuals possess the capability to organize and execute courses of action required to attain particular outcomes (Bandura, 1986). Since Bandura introduced the construct of self-efficacy into the psychological literature, it has become one of the most extensively studied psychological constructs. Researchers have explored its role across various domains, including education, health, career, and well-being, by examining self-efficacy among individuals from different developmental levels and cultural backgrounds (Schunk & Pajares, 2002).

In line with the development of self-efficacy research, several instruments have been developed to measure self-efficacy, including the Generalized Self-Efficacy Scale (GSE; Schwarzer and Jerusalem 1995), the General Self-Efficacy Scale (GSES; Sherer et al. 1982), and the General Self-Efficacy Scale 12 (GSES-12; Bosscher and Smit 1998). The Generalized Self-Efficacy Scale (GSE), developed by Schwarzer and Jerusalem (1995), is used to measure the general construct of self-efficacy. The GSE assesses general perceptions of self-efficacy with the aim of predicting how individuals cope with everyday problems and adapt after experiencing various stressful life events.

The GSE was initially developed by Matthias and Jerusalem in Germany in 1981. It originally consisted of 20 items designed to measure self-efficacy as a broad personality disposition that could be assessed individually or collectively. In 1995, Schwarzer and Jerusalem (1995) modified the instrument into a 10-item, one-factor (unidimensional) measure. The instrument reflects optimistic self-confidence, namely the belief that individuals can perform new or difficult tasks and overcome adversity. The GSE is currently available in 33 languages, including Indonesian. Across studies conducted in many countries, the GSE has demonstrated good

© GamaJOP 2026. This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

internal consistency, with Cronbach's alpha values ranging from 0.76 to 0.90 and an average reliability coefficient of approximately 0.80 (Schwarzer & Jerusalem, 1995).

Versions of the GSE in various languages have been examined using factor analysis and have provided evidence supporting construct validity. Nevertheless, despite its widespread acceptance, adapting and translating the GSE into different cultural contexts, such as Indonesia, presents specific challenges. One such challenge is ensuring that the GSE is not only theoretically and structurally valid but also fair when measuring self-efficacy across different groups. This is important because cultural, educational, gender, and social differences may influence how individuals evaluate their own capabilities. Previous studies have shown that self-efficacy is influenced by cultural and social factors, including gender differences (Britner & Pajares, 2001; Mulu, 2023). Therefore, it is important to examine whether the Indonesian version of the GSE exhibits differential item functioning (DIF).

Studies of DIF have attracted considerable attention in psychological and educational research. DIF analysis is used to support measurement fairness and strengthen test validity (Ibrahim, 2018). DIF refers to a phenomenon in which a test item functions differently across groups even when individuals from those groups have the same level of the underlying ability. Examining whether an item exhibits DIF can help researchers better understand item and group performance, thereby supporting fairer test use (Penfield & Camilli, 2006).

Several methods are currently available for detecting DIF. DIF detection methods can be classified into two broad philosophical perspectives: statistical methods (e.g., factor-analysis-based approaches) and measurement models (e.g., the Rasch model), each of which is based on different assumptions (Hayat et al., 2023). Previous studies have applied approaches such as Item Response Theory (Stark et al., 2006), the Rasch model (Suseno et al., 2022), and the Mediated MIMIC model (Cheng et al., 2016; Wahyudi, 2019; W.-C. Wang et al., 2009). The MIMIC model is a special case of confirmatory factor analysis (CFA) that incorporates covariates and includes observed variables serving as multiple indicators and multiple causes of a latent variable (Bollen, 1989).

Several studies have examined DIF in the GSE, which is widely used to measure self-efficacy. Bonsaksen et al. (2013), for example, used the Rasch model in a sample of individuals with morbid obesity. Their findings showed significant DIF associated with age and educational level. Several items were easier to endorse for individuals younger than 40 years or those with higher educational attainment. In addition, Holter et al. (2025) found that one item (Item 3) exhibited DIF in relation to gender responses. Women with low self-efficacy were more likely to endorse this item, whereas men were more likely to endorse it when their self-efficacy was at a moderate to high level.

Other studies of gender differences have also reported relevant findings. Robinson et al. (2022), for example, found that women had lower levels of self-efficacy than men, which affected their motivation and academic achievement in science, technology, engineering, and mathematics (STEM). Similar findings were reported by Chan (2022) in China, where women showed lower levels of self-efficacy, interest, and academic and career aspirations in STEM fields than men.

DIF analysis is essential for ensuring that self-efficacy measures such as the GSE can be used fairly across different groups in the Indonesian context, which is characterized by sociocultural diversity, varied educational backgrounds, and gender differences. DIF analysis is necessary as part of a comprehensive instrument-validation process to examine measurement invariance, particularly whether GSE items measure the same construct across different groups. DIF analysis can help identify potentially biased items so that test scores can be interpreted fairly and accurately (Gómez-Benito et al., 2018; Hagquist, 2019; Martinková et al., 2017). Gender, for example, may create differences in how men and women assess their levels of self-efficacy, potentially reflecting social norms and societal expectations regarding gender roles (Kutuk, 2025).

This study had two objectives. First, it examined the factor structure of the Indonesian version of the GSE using CFA. Second, it analyzed DIF using the Mediated MIMIC model, with gender as the covariate. DIF analysis should be conducted as part of psychometric evaluation to determine whether the Indonesian version of the GSE functions fairly across different groups. The findings of this study are expected to provide insight into the extent to which the Indonesian version of the GSE is valid and fair for measuring self-efficacy in the Indonesian population. In addition, the findings are expected to contribute to the development of psychometric instruments that are more sensitive to gender, sociocultural, and educational factors, particularly in the Indonesian context.

Methods

Participants

The participants were 534 undergraduate students (102 males and 432 females) enrolled at various universities in Indonesia. Their ages ranged from 17 to 25 years (M = 20.06, SD = 1.34). A non-probability sampling method with a convenience sampling technique was used.

Ethical Approval and Research Procedure

This study was conducted after receiving ethical approval from the Research Ethics Committee of Universitas Bosowa, approval number: 003b/1.02/DRIPM-UNIBOS/2025 dated January 3, 2025. All participants were informed about the purpose of the study, the voluntary nature of participation, data confidentiality, and their right to withdraw at any time without consequences. Data were collected through an online questionnaire administered via Google Forms. All participants provided informed consent before completing the research instrument. All data were analyzed anonymously and used solely for research purposes.

Instrument: Indonesian Version of the Generalized Self-Efficacy Scale (GSE)

This study used the GSE developed by Schwarzer and Jerusalem (1995). The GSE is a unidimensional instrument consisting of 10 items rated on a four-point Likert scale, ranging from Disagree = 1 to Strongly Agree = 4. The instrument reflects optimistic self-confidence, indicating the belief that individuals can perform new or difficult tasks and overcome related difficulties. The GSE is available in 33 languages, including an Indonesian version. The Indonesian version follows the scoring system of the original instrument, using a four-point Likert scale, with total scores ranging from 10 to 40.

Data Analysis: Mediated Multiple Indicators Multiple Causes (MIMIC) Model

The MIMIC model can incorporate observed covariates, such as gender and ethnicity, to examine their effects on latent variables and item responses, while simultaneously examining how these variables influence item responses. One advantage of the MIMIC model is its ability to control for covariates (e.g., gender), making it a flexible and widely used method for detecting DIF in psychometric studies (Finch, 2005; W.-C. Wang et al., 2009; Woods & Grimm, 2011).

The MIMIC model is a special case of structural equation modeling (SEM) in which multiple indicators reflect an underlying latent variable or factor and multiple causes (predictors) influence that latent variable or factor. When the covariance structure is analyzed, the MIMIC model can be represented as follows:

$$\eta = \Gamma X + \zeta \quad (1)$$
$$Y = \lambda_y \eta + \varepsilon \quad (2)$$
$$X \equiv \xi \quad (3)$$

In these equations, multiple endogenous indicators (y's) are used to measure endogenous latent variables (η's). No causal effects are specified among the η's; instead, covariances/correlations are specified, and η is influenced by exogenous indicators (x's), which are assumed to be perfect measures of exogenous latent variables (ξ's). The symbol $\equiv$ indicates identity between x and ξ by fixing the factor loading to 1.0 (i.e., $\Lambda_x = 1$) and the measurement error to 0 (i.e., $\Theta_\delta = 0$). When means and covariances are analyzed, the MIMIC model can be represented as follows:

$$\eta = \Gamma X + \zeta \quad (4)$$
$$Y = \upsilon_y + \lambda_y \eta + \varepsilon \quad (5)$$
$$X \equiv \xi \quad (6)$$

Here, υ_y represents the vector of means/intercepts of the endogenous indicators y. In the equations above, no factor intercept is included because the factor mean/intercept in the single-group model is fixed at 0 for model identification (J. Wang & Wang, 2020).

Data Analysis Procedure

Data were analyzed using CFA and the Mediated MIMIC model. To obtain information about the psychometric characteristics of each subscale, the following analyses were reported: (1) goodness-of-fit testing for the CFA and Mediated MIMIC models; (2) significance testing of item parameters; (3) significance testing of the regression coefficient from the covariate, namely gender, to the factor; (4) reporting of correlations among factors estimated from the model (latent correlations); and (5) comparison of the score distributions from the CFA and Mediated MIMIC models. In the Mediated MIMIC analysis, gender was used as a covariate and coded dichotomously as male = 1 and female = 0.

Results

Descriptive Statistics

Before conducting CFA, descriptive statistics were examined by reporting the mean, standard deviation, and skewness of each item in the Indonesian version of the GSE. The normality assumption was evaluated using the skewness values for each item, with an acceptable range of ± 1 (B. Muthén & Kaplan, 1992). The results indicated that the normality assumption was not fully met because some items had skewness values outside the specified range (see Table 2).

Model Testing

CFA examines the relationships between observed variables and continuous latent variables. CFA is one of the most commonly used approaches for evaluating the construct validity of an instrument (Brown, 2015). Using CFA, researchers can examine the extent to which all items in an instrument measure or provide information about a single intended construct (Umar & Nisa, 2020). Several goodness-of-fit indices were considered in the CFA, including the root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker–Lewis index (TLI), and standardized root mean square residual (SRMR). A model was considered to have acceptable fit when RMSEA < 0.08, CFI and TLI > 0.90, and SRMR < 0.08 (Browne & Cudeck, 1993; Hu & Bentler, 1999).

The first stage involved conducting CFA on the Indonesian version of the GSE. The analysis was performed using Mplus version 8.4 (L. K. Muthén & Muthén, 2020) with maximum likelihood estimation with standard errors and mean-adjusted chi-square test statistics that are robust to non-normality (ESTIMATOR = MLM). Initial testing of the one-factor (unidimensional) model of the Indonesian GSE yielded RMSEA = .068 (90% CI = .056, .082), CFI = .937, TLI = .919, and SRMR = .045. These results indicate that the model was supported by the data (Browne & Cudeck, 1993). Thus, the 10 items of the Indonesian GSE measured a single factor, namely self-efficacy. In addition, all 10 items showed statistically significant factor loadings.

After the one-factor model was found to fit the data, gender was included in the model. With gender included, the one-factor (unidimensional) model was specified as a

Table 1
Indonesian Version of the General Self-Efficacy Scale

No	Indonesian Version of the General Self-Efficacy Scale
1.	*Menyelesaikan masalah yang sulit, selalu berhasil bagi saya, jika saya mencobanya*
2.	*Jika seseorang menghalangi tujuan saya, saya akan mencari cara dan sarana untuk melanjutkan*
3.	*Saya tidak mengalami kesulitan dalam menjalankan maksud dan tujuan saya.*
4.	*Dalam situasi yang tidak terduga, saya selalu tahu bagaimana saya harus bersikap.*
5.	*Jika saya dihadapkan pada sesuatu yang baru, saya tahu bagaimana saya dapat menghadapinya.*
6.	*Untuk setiap masalah, saya punya solusinya.*
7.	*Saya dapat menghadapi kesulitan dengan tenang, karena saya selalu dapat mengandalkan kemampuan saya.*
8.	*Jika saya menghadapi suatu masalah, saya biasanya mempunyai banyak ide untuk mengatasinya.*
9.	*Juga dalam kejadian yang tidak terduga, saya pikir saya akan mampu menghadapinya dengan baik.*
10.	*Apapun yang terjadi, saya akan siap menghadapinya.*

Table 2
Descriptive Statistics

Item	Mean	Std. Deviation	Skewness
1.	3.380	0.775	−1.042
2.	3.479	0.732	−1.342
3.	2.433	0.987	−0.032
4.	3.047	0.833	−0.518
5.	2.820	0.834	−0.178
6.	2.901	0.824	−0.339
7.	2.831	0.909	−0.292
8.	2.805	0.859	−0.095
9.	2.828	0.806	−0.066
10.	3.337	0.796	−0.971

Mediated MIMIC model. The analysis showed that the Mediated MIMIC model fit the data, with RMSEA = 0.062 (90% CI = 0.050, 0.074), CFI = 0.939, TLI = 0.922, and SRMR = 0.043 (Browne & Cudeck, 1993). Accordingly, the effect of gender differences on the Indonesian version of the GSE could be interpreted. After establishing that both the one-factor CFA model and the Mediated MIMIC model of the Indonesian GSE fit the data, the path diagrams of the fitted models and the item-level parameter estimates were reported.

Table 3 presents the parameter estimates for each item in both the CFA and Mediated MIMIC models. Table 4 presents the regression coefficients from the Mediated MIMIC analysis.

As shown in Table 3, all items of the Indonesian version of the GSE demonstrated statistically significant factor loadings in both the CFA and Mediated MIMIC models. Item performance was evaluated based on two criteria: positive factor loadings and statistically significant z-values greater than 1.96 ($p < .05$). As presented in Table 3, all items met these criteria, indicating that each item significantly contributed to the measurement of the underlying self-efficacy construct (Umar & Nisa, 2020).

The Mediated MIMIC analysis indicated gender-related DIF in the Indonesian version of the GSE. This result was reflected in statistically significant z-values greater than 1.96 (Table 4). Further analysis was then conducted to identify the item exhibiting DIF. The Mediated MIMIC analysis showed that Item 3 ("I have no difficulty in carrying out my intentions and goals") exhibited gender-related DIF through both direct and indirect effects (see Figure 2). The z-value for Item 3 was greater than 1.96 (see Table 4). This finding indicates that men and women responded differently to Item 3. The result may reflect differences in gender roles within the Indonesian sociocultural context that influence how individuals evaluate their own capabilities.

Distribution of CFA and Mediated MIMIC Model Scores

The score distributions for the CFA and Mediated MIMIC models were examined. For the CFA model (Figure 3), the mean was 0.000 (SD = 0.268), with a skewness of −0.067. For the Mediated MIMIC model (Figure 4), the mean was 0.074 (SD = 0.947), with a skewness of −0.065. Although an independent variable was added to the model (the Mediated MIMIC model), the mean and skewness values were relatively similar across the two models, whereas the standard deviations indicated differences in the dispersion of the factor scores.

Discussion

This study aimed to analyze the factor structure and identify DIF in the Indonesian version of the GSE using CFA and the Mediated MIMIC model. These approaches are frequently used to evaluate the psychometric properties of instruments because they allow researchers to assess model fit and the factor structure of latent constructs (Brown, 2015). The findings provide important insights

Figure 1
Fitted model of the Indonesian version of the GSE

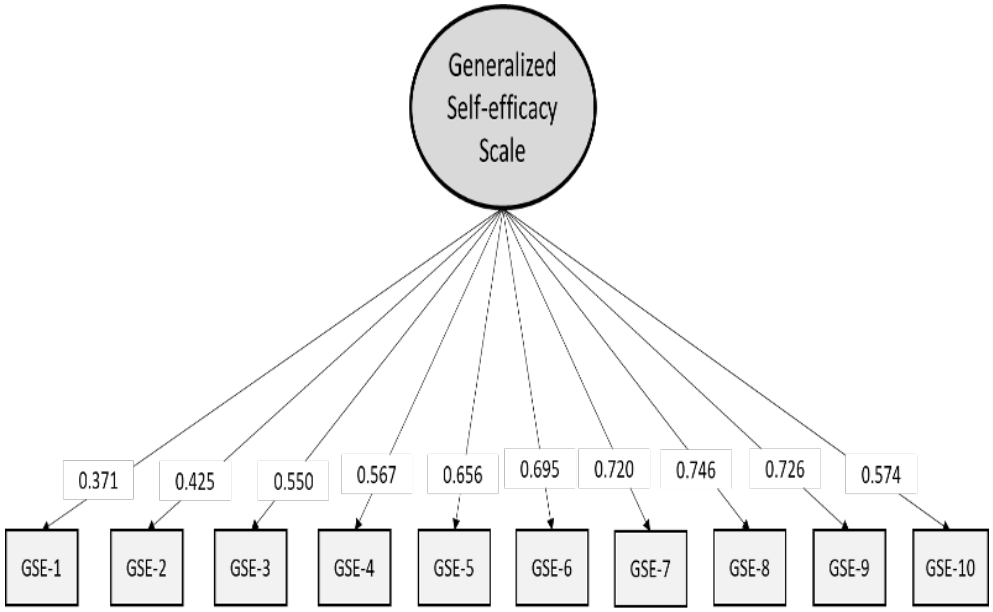


Figure 2
Fitted Mediated MIMIC model of the Indonesian version of the GSE

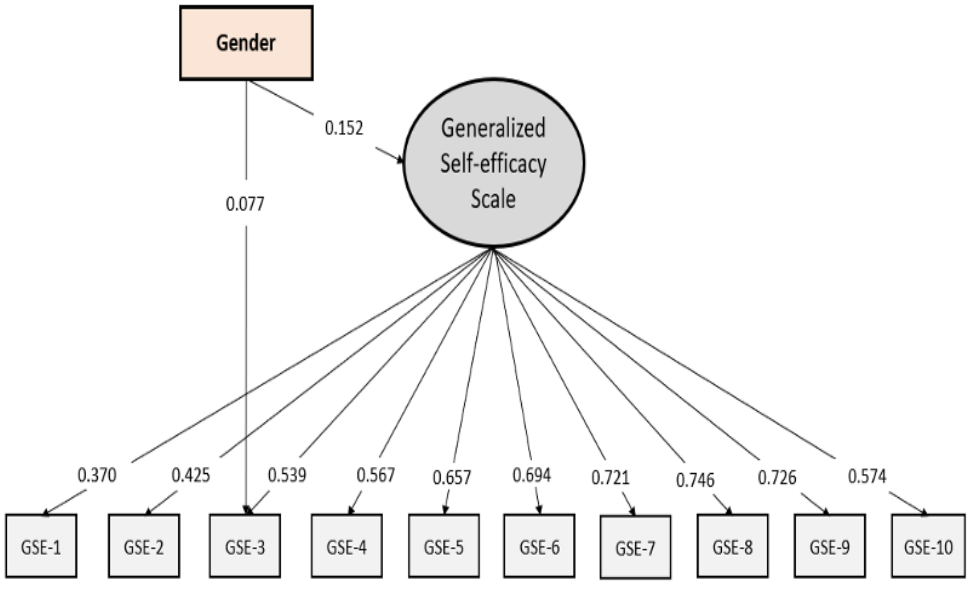


Table 3
Item Parameters for the CFA and Mediated MIMIC Models

	CFA			Mediated MIMIC model			Interpretation
Item	Loading	S.E.	z-value	Loading	S.E.	z-value	
1	0.371	0.041	8.935	0.370	0.042	8.912	Significant loading
2	0.425	0.037	11.417	0.425	0.037	11.404	Significant loading
3	0.550	0.036	15.151	0.539	0.037	14.534	Significant loading
4	0.567	0.038	14.883	0.567	0.038	14.912	Significant loading
5	0.656	0.030	21.838	0.657	0.030	21.808	Significant loading
6	0.695	0.026	26.782	0.694	0.026	26.731	Significant loading
7	0.720	0.025	28.499	0.721	0.025	28.636	Significant loading
8	0.746	0.023	32.787	0.746	0.023	32.924	Significant loading
9	0.726	0.028	25.725	0.726	0.028	25.731	Significant loading
10	0.574	0.035	16.405	0.574	0.035	16.380	Significant loading

Table 4
Regression Coefficients for the MIMIC Model

Path	Beta	S.E.	z-value
Direct effects			
Gender → General Self-Efficacy	0.152	0.046	3.315
Gender → Item 3	0.077	0.036	2.139
Indirect effect			
Gender → General Self-Efficacy → Item 3	0.082	0.025	3.230

Figure 3
Distribution of CFA scores

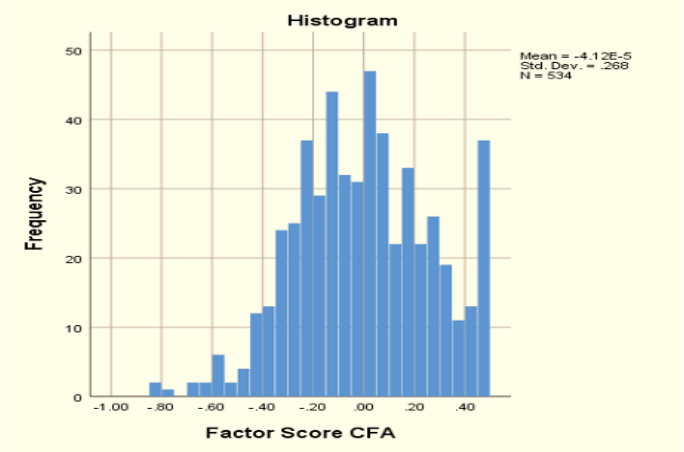
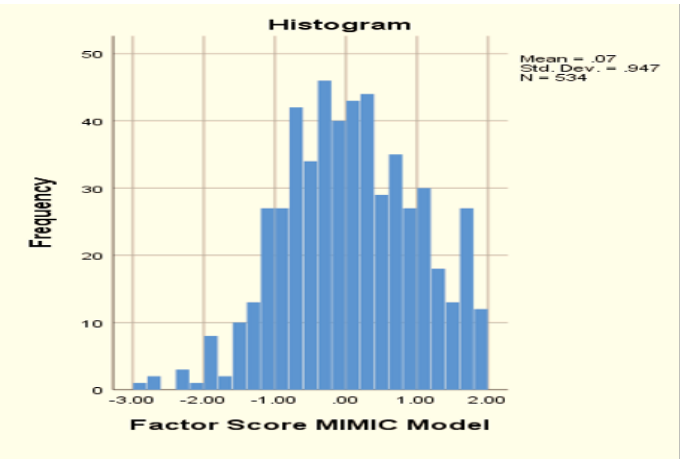


Figure 4
Distribution of Mediated MIMIC model



into the validity and fairness of the instrument in the Indonesian context, particularly for assessing self-efficacy among university students. CFA showed that the one-factor (unidimensional) model of the Indonesian GSE fit the data. In addition, all items showed statistically significant factor loadings, ranging from 0.370 (Item 1) to 0.746 (Item 8). These findings support the results reported for the original scale by Schwarzer and Jerusalem (1995).

The Mediated MIMIC analysis further indicated a gender effect on the Indonesian version of the GSE. The presence of gender-related DIF indicates potential measurement bias that should be considered when comparing scores across gender groups. In an international context, previous Rasch analyses of the GSE have identified DIF in specific items according to age, education, and employment status (Bonsaksen et al., 2013). Thus, the present findings reinforce the importance of evaluating DIF when adapting psychometric instruments across groups in Indonesia and highlight the need for revision or corrective measures to ensure fair and reliable GSE measurement.

Several previous studies are consistent with the findings of the present study. Britner and Pajares (2001) found that men tended to report higher levels of self-efficacy than women in educational contexts. This finding reflects differences in how men and women evaluate their abilities to cope with academic challenges. In developing the GSE, Schwarzer and Jerusalem (1995) reported that although the instrument did not show substantial DIF across cultures, differences between men and women might emerge in responses to items influenced by sociocultural and educational backgrounds. Mulu (2023) also found significant differences in self-efficacy between male and female students, with male students tending to report higher self-efficacy than female students.

Further analysis showed that the item exhibiting DIF was Item 3 ("I have no difficulty in carrying out my intentions and goals") (see Table 4). Item 3 showed DIF through both direct and indirect effects, indicating differences in how men and women responded to the item. This finding is consistent with Holter et al. (2025), who identified gender-related DIF in Item 3 of the GSE. Women with low self-efficacy were more likely to endorse this item, whereas men were more likely to endorse it when their self-efficacy was at a moderate to high level. In addition, Husain (2014) found that male students demonstrated higher self-efficacy than female students. These findings may reflect differences in gender roles within the Indonesian sociocultural context that influence how individuals evaluate their own capabilities, which may be reflected in responses to Item 3.

Another aspect examined in this study was the comparison of the score distributions from the CFA and Mediated MIMIC models in terms of the mean, standard deviation, and skewness. The mean values were nearly identical (0 vs. 0.074), indicating no meaningful shift in the center of the distribution. The standard deviation was larger for the Mediated MIMIC model (0.947 vs. 0.268), indicating that the inclusion of the independent variable (gender) increased score dispersion without substantially changing the shape of the distribution. Similarly, the skewness values (−0.065 vs. −0.067) were both close to zero, indicating approximately symmetric distributions in both models. The CFA and Mediated MIMIC models therefore produced factor-score distributions that were nearly identical in terms of central tendency and shape, with differences primarily in dispersion (SD). This indicates that the robustness of the CFA was maintained despite the expanded model. A similar study by Filho et al. (2014) also reported comparable latent-trait distributions across models.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, only gender was considered as a covariate in the DIF analysis. Consequently, the study cannot explain potential differential item functioning associated with other characteristics, such as age, educational background, culture, geographic region, or socioeconomic conditions. Testing a broader range of covariates is necessary to obtain more comprehensive evidence of measurement invariance for the Indonesian version of the GSE.

The second limitation concerns the gender imbalance in the sample (102 males and 432 females), representing a ratio of approximately 1:4. Although this ratio is still considered acceptable for DIF analysis (Penfield & Camilli, 2006), it may reduce estimation precision and the sensitivity of DIF detection in the smaller group, as well as limit the generalizability of the findings (Zumbo, 1999, 2007). Third, the study used convenience sampling. This technique was selected because of the ease of access to participants, but it has a fundamental limitation in terms of representativeness (Etikan et al., 2016). Therefore, the findings cannot be fully generalized to the broader population.

Conclusion

This study provides empirical evidence regarding the factor structure of the Indonesian version of the GSE using CFA while incorporating DIF detection through the Mediated MIMIC model. The findings showed that the Indonesian GSE had an acceptable one-factor (unidimensional) model fit. In addition, all items showed statistically significant factor loadings as indicators of self-efficacy. However, despite the adequate psychometric performance of the instrument, gender-related DIF was detected. The item identified as exhibiting DIF was Item 3. The presence of DIF in the Indonesian version of the GSE should therefore be considered when the instrument is used.

Implications

The findings have important implications for the use of the Indonesian version of the GSE in research and psychological assessment practice. Support for the one-factor structure indicates that the 10 items generally represent the construct of general self-efficacy among Indonesian

university students. However, the presence of gender-related DIF in Item 3 indicates that caution is warranted when GSE scores are used to compare male and female respondents, as this item may not function equivalently across gender groups.

Recommendations

Future studies are recommended to re-examine the Indonesian version of the GSE using larger and more diverse samples with a more balanced gender distribution. DIF analyses should also be expanded by considering other covariates, such as age, educational level, geographic region, cultural background, and socioeconomic characteristics. Item 3, in particular, should be evaluated using both quantitative and qualitative approaches, such as cognitive interviewing, to determine whether the observed DIF arises from linguistic formulation, semantic interpretation, or differences in social experiences across groups. Cross-validation testing and the use of alternative DIF methods are also needed to determine whether the findings for Item 3 are consistent and replicable.

Declarations

Acknowledgment

The authors sincerely acknowledge all participants who voluntarily contributed to this study. Appreciation is also extended to the academic colleagues and institutions that facilitated the data collection process and provided support during the completion of this research. Their contributions were valuable to the successful conduct of the study.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

RSB: Conceptualization, methodology, formal analysis, data interpretation, visualization, writing—original draft, writing—review and editing, and project administration. B: Conceptualization, theoretical framework development, and writing—review and editing. ADP: Data collection, data curation, writing—review and editing.

Declaration of Generative AI in Scientific Writing

The authors used ChatGPT to review the manuscript for compliance with academic writing standards and to check for typographical errors and inconsistencies in wording and sentence construction. The authors reviewed and verified the final manuscript and take full responsibility for its content.

Orcid ID

Rahmat S. Bintang: https://orcid.org/0009-0000-3233-5444
Burhan: https://orcid.org/0009-0007-3293-8692
Anhar Dana Putra: https://orcid.org/0000-0001-5818-0129

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Bonsaksen, T., Kottorp, A., Gay, C., Fagermoen, M. S., & Lerdal, A. (2013). Rasch analysis of the general self-efficacy scale in a sample of persons with morbid obesity. *Health and Quality of Life Outcomes*, *11*, Article 202. https://doi.org/10.1186/1477-7525-11-202

Bosscher, R. J., & Smit, J. H. (1998). Confirmatory factor analysis of the general self-efficacy scale. *Behaviour Research and Therapy*, *36*(3), 339–343. https://doi.org/10.1016/S0005-7967(98)00025-4

Britner, S., & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering*, *7*(4), 15. https://doi.org/10.1615/jwomenminorscieneng.v7.i4.10

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Chan, R. C. H. (2022). A social cognitive perspective on gender disparities in self-efficacy, interest, and aspirations in science, technology, engineering and mathematics (STEM): The influence of cultural and gender norms. *International Journal of STEM Education*, *9*, Article 47. https://doi.org/10.1186/s40594-022-00352-0

Cheng, Y., Shao, C., & Lathrop, Q. N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, *76*(1), 43–63. https://doi.org/10.1177/0013164415576187

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11

Filho, N. H., Machado, W. D. L., & Damásio, B. F. (2014). Effects of statistical models and item difficulties on making trait-level inferences: A simulation study. *Psychology/Psicologia: Reflexão e Crítica*, *27*(4), 670–678. https://doi.org/10.1590/1678-7153.201427407

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with mantel-haenszel, sibtest, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. https://doi.org/10.1177/0146621605275728

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104–109. https://doi.org/10.7334/psicothema2017.183

Hagquist, C. (2019). Explaining differential item functioning focusing on the crucial role of external information: An example from the measurement of adolescent mental health. *BMC Medical Research Methodology*, *19*, Article 216. https://doi.org/10.1186/s12874-019-0828-3

Hayat, B., Rahayu, W., Putra, M. D. K., Sarifah, I., Deviana, T., Puri, V. G. S., & Isa, K. (2023). Metacognitive skills assessment in research-proposal writing (MSARPW) in the indonesian university context: Scale development and validation using multidimensional item response models. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, *12*(1), 31–47. https://doi.org/10.15408/jp3i.v12i1.31679

Holter, M., Avian, A., Weger, M., Strini, S., Michelitsch, M., Winkler, V., Kloft, A. M., Groß, J., Falb, T., Gabriel, M., Großpötzl, M., Wedrich, A., & Berghold, A. (2025). Uncovering potential interviewer-related biases in self-efficacy assessment: A study among chronic disease patients. *BMC Psychology*, *13*(1), Article 299. https://doi.org/10.1186/s40359-025-02579-2

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Husain, U. (2014). *Relationship between self-efficacy and academic motivation*. https://doi.org/10.15242/icehm.ed1214132

Ibrahim, A. W. (2018). Differential item functioning: The state of the art. *Jigawa Journal of Multidisciplinary Studies*, *1*(1), 37–50.

Kutuk, G. (2025). Understanding gender stereotypes in the context of foreign language learning through the lens of social cognitive theory. *TESOL Journal*, *59*, 1081–1105. https://doi.org/10.1002/tesq.3267

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), Article rm2. https://doi.org/10.1187/cbe.16-10-0307

Mulu, M. (2023). Self-efficacy: A view from junior high school students and its gender interaction. *Journal of Educational Science and Technology*, *9*(2), 150. https://doi.org/10.26858/est.v9i1.48373

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 19–30. https://doi.org/10.1111/j.2044-8317.1992.tb00975.x

Muthén, L. K., & Muthén, B. O. (2020). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In *Handbook of statistics*. Elsevier Science & Technology. https://doi.org/10.1016/S0169-7161(06)26005-X

Robinson, K. A., Perez, T., White-Levatich, A., & Linnenbrink-Garcia, L. (2022). Gender differences and roles of two science self-efficacy beliefs in predicting post-college outcomes. *The Journal of Experimental Education*, *90*(2), 344–363. https://doi.org/10.1080/00220973.2020.1808944

Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 15–31). Academic Press. https://doi.org/10.1016/B978-012750053-9/50003-6

Schwarzer, R., & Jerusalem, M. (1995). General self-efficacy scale (general self-efficacy) [Database record]. https://doi.org/10.1037/t00393-000

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports*, *51*(2), 663–671. https://doi.org/10.2466/pr0.1982.51.2.663

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Suseno, M., Hayat, B., Putra, M. D. K., Bien, J. K., Rachmawati, R., & Hartanto, H. (2022). A differential item functioning (DIF) analysis of the mobile phone problem use scale in indonesian schools with and without smartphone banned policy. *Cogent Psychology*, *9*(1), Article 2137306. https://doi.org/10.1080/23311908.2022.2137306

Umar, J., & Nisa, Y. F. (2020). Uji validitas dengan CFA dan pelaporannya. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, *9*(2), 1–11. https://doi.org/10.15408/jp3i.v9i2.16964

Wahyudi, H. (2019). Pengaplikasian multiple indicator multiple causes (MIMIC) model dalam mendeteksi differential item functioning (DIF) pada alat ukur social quality of life. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, *8*(1), 25–36. https://doi.org/10.15408/jp3i.v8i1.12581

Wang, J., & Wang, X. (2020). *Structural equation modeling: Applications using mplus* (2nd ed.). John Wiley & Sons.

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*(5), 713–731. https://doi.org/10.1177/0013164409332228

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361. https://doi.org/10.1177/0146621611405984

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Directorate of Human Resources Research; Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832