TPM®

MULTIDIMENSIONALITY OF STUDENT ENGAGEMENT CONSTRUCT: THE EXPLORATORY AND CONFIRMATORY ITEM RESPONSE MODEL

ALI RIDHO
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG, INDONESIA

This study demonstrates the use of the multidimensional item response theory (MIRT) to investigate the internal structure of a construct exploratively and confirmatively. Based on data from 657 Islamic university students (65% female) spread across Indonesia, MIRT was used to examine the factor structure of the items measuring student engagement. The MIRT results supported the multidimensional structure of the scale. Most notably, the comparison of the investigated models supported the within-item multidimensional structure in which almost all items fit 3-factor loadings among all measured domains (cognitive, behavioral, and social). Furthermore, vector depictions of the items in a 3-dimensional space are offered to give the reader a vivid picture of their multidimensionality. The paper ends with an overview of MIRT in scale development and dimensionality assessment to didactically enhance readers' awareness of its usefulness as a psychometric tool.

Keywords: Student engagement; Confirmatory factor analysis; Dimensionality assessment; Exploratory factor analysis; Multidimensional item response theory (MIRT).

Correspondence concerning this article should be addressed to Ali Ridho, Faculty of Psychology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Jalan Gajayana 50, Malang 65144, East Java, Indonesia. Email: aliridho@uin-malang.ac.id

Empirical research on the multidimensionality of student engagement is an essential endeavor in psychometrics and higher education. Student engagement has attracted considerable attention due to its strong association with academic success, motivation, critical thinking skills, and students' sense of belonging in the educational environment (Salmela-Aro et al., 2021). Diverse research perspectives reinforce it. Yang et al. (2023) highlighted the pivotal role of engagement in mitigating dropout rates and fostering academic achievement, especially under remote learning conditions during the pandemic. Siu et al. (2023) identified psychological capital and study engagement as mediators in the relationship between social support and student outcomes, thus underscoring the importance of engagement in academic and behavioral performance. Nkomo et al. (2021) explored the nuances of student engagement in digital learning environments, stressing its impact on developing critical thinking skills and student satisfaction. Complementing these findings, Li and Xue (2023) conducted a meta-analysis revealing promoting and hindering student engagement factors, such as the teacher-student relationship and teaching methods, affecting students' learning participation and success in various educational contexts. These collective insights highlight the multifaceted nature of engagement, its profound impact on multiple aspects of the educational experience, and the importance of a supportive learning environment in fostering effective engagement. Consequently, as researchers, educators, and policymakers, understanding the complex nature of student engagement is paramount.

Traditionally, student engagement has been conceptualized and measured as a unidimensional construct, which ignores the complex interplay of its dimensions. However, recent research has challenged this simplistic view, recognizing that engagement is multifaceted. For example, recently, researchers have improved

TPM Vol. 31, No. 2, June 2024 – 239-261 – doi:10.4473/TPM31.2.7 – © 2024 Cises

educationalists' understanding of student engagement as multidimensional (Inman et al., 2020; Wong & Liem, 2022; Zhang & McNamara, 2018). Other empirical studies have also proven the multidimensional structure of student engagement, consisting of affective, behavioral, and cognitive (Ben-Eliyahu et al., 2018; Bond et al., 2020; Groccia, 2018), as well as social dimensions (Fredricks et al., 2019; Rimm-Kaufman et al., 2015; Wang et al., 2016; Wang & Hofkens, 2020). Additionally, a synthesis through a scoping review of 2010-2020 publications on student engagement by Salmela-Aro et al. (2021) confirmed that student engagement is a multidimensional construct.

This growing consensus encourages exploration of the multidimensionality of student engagement, prompting empirical investigations to reveal the specific dimensions that contribute to the overall student engagement experience. While these empirical studies have demonstrated the multidimensional nature of the student engagement construct, they have yet to apply the multidimensional item response theory (MIRT) method as Carlucci et al. (2023) conducted. Carlucci et al. utilized a sample of 3338 Italians to assess the multidimensional nature of anxiety using the state-trait inventory for cognitive and somatic anxiety (STICSA). By comparing unidimensional, 2-factor, and bifactor models, the research demonstrated that a bifactor model most accurately captured the nuances of state and trait anxiety, underscoring the construct multidimensionality. Furthermore, attention to the multidimensionality of student engagement needs to be prioritized, because this affects the scoring and the interpretation of scores.

The score becomes the basis for concluding a test taker on the attribute of concern and is the foundation for estimating candidates' ability, proficiency level, or latent trait level. Meanwhile, the estimated latent attribute of proficiency depends on the measurement model and refers to the interpretation of the score influenced by the internal structure of the model chosen. Thus, the model selected affected the validity related to the scoring interpretation (AERA et al., 2014). Scores will be useful and produce meaningful information if they have good psychometric properties and are valid and reliable. The most important consideration is the validity of score interpretation, which is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). While scores and their meanings depend on the structure of the construct (Childs & Oppler, 2000) and ignoring the better model may result in misleading conclusions (Lee et al., 2019), this implies that investigating the structure of the construct component dimensions is crucial to support the validity of the score interpretations.

Of the sources of validity evidence, internal structure validity is a prerequisite in developing arguments to support the interpretation and use of scores (Rios & Wells, 2014). Meanwhile, test scores depend on the internal structure of the instrument. Internal structure validity addresses the extent to which the relationships between items and latent dimensions align with theoretical expectations. Various factor analysis models are available to investigate an instrument factor structure. In addition to exploratory factor analysis (EFA; Watson, 2017), confirmatory factor analysis (CFA) has been widely used as a model-based approach to assess the factor structure of an instrument (Mikkonen et al., 2022). Alternatively, item response theory (IRT), which represents a broad class of statistical models, can also be used to help conduct item analysis, scale development, and scoring (Embretson, 1996; Embretson & Reise, 2000; Liu et al., 2018). While at its inception, traditional IRT was limited to using the principle of unidimensionality, that is, only one latent attribute (θ) underlying test takers' responses, recent advances in the field have led to the development of multidimensional IRT models (Reckase, 2009) that can be used for multidimensional instruments where the latent attributes are as many as m dimensions ($\theta_1, \theta_2, \ldots, \theta_m$). For example, $m = 3$ in the engagement construct, θ_1 = cognitive engagement, θ_2 = behavioral engagement, and θ_3 = social engagement.

The classical test theory (CTT) is limited in modeling multidimensional constructs, whereas MIRT effectively extends from unidimensional to multidimensional models. MIRT capacity for multidimensionality is evident in its application to complex analytic geometry tests, whereas MIRT surpasses CTT in analyzing a wide range of competencies (Kruglova et al., 2021). Moreover, integrating MIRT with graphical models provides a robust measurement of multidimensional traits, outperforming CTT (Y. Chen et al., 2018). The clinical assessment further underscores these advantages, while MIRT multidimensional models yield more nuanced and accurate representations than CTT (Thomas, 2019).

The application of MIRT to assess internal factor structure in psychological research needs to be more developed. Recognizing this gap, this study aims to demonstrate the utility of MIRT modeling in unraveling the complexities of instruments' factor structures, exemplified by the studies conducted by C.-Y. Chen et al. (2018) and Carlucci et al. (2023). Chen and colleagues used MIRT to analyze the ages and stages questionnaires (ASQ-3, 3rd ed.), revealing its internal structure as multidimensional rather than unidimensional, encompassing interrelated domains such as fine motor, gross motor, communication, problem-solving, and personal-social skills. This finding highlights the importance of using MIRT to capture the complex, interconnected nature of constructs like child development, which traditional unidimensional approaches might oversimplify.

Similarly, Carlucci et al. (2023) applied MIRT to the state-trait inventory for cognitive and somatic anxiety (STICSA), addressing its multidimensional nature that encompasses state and trait anxiety with both cognitive and somatic components. Their study found that both bifactor and 2-correlated dimensions models fit the STICSA scales well, with the bifactor model showing better-fit indices. However, the multidimensional model provides more precision in estimating latent anxiety states. This underscores the strength of MIRT in providing a more accurate and comprehensive assessment of complex psychological constructs. Moreover, the authors extended the application of MIRT beyond developmental and anxiety measures to student engagement assessment. Student engagement, which can manifest cognitively, behaviorally, socially, or as an interaction of these domains, is inherently multidimensional. MIRT-based scores offer a more comprehensive understanding of student engagement by capturing its multifaceted nature, allowing educators and researchers to tailor interventions more effectively and understand the diverse ways students interact with their learning environments. The findings from C.-Y. Chen et al. (2018) and Carlucci et al. (2023) collectively reinforce the advantage of MIRT in providing nuanced insights into complex constructs, ensuring a more accurate, reliable, and comprehensive assessment essential in various fields of psychological research and education.

The scope of this empirical research is to provide a comprehensive understanding of the multidimensional nature of student engagement. Its findings will yield valuable insights that can inform educational practice, empower students to thrive academically, personally, and socially, and ultimately advance the field of psychometrics and higher education research. Moreover this study should inspire further investigation and discussion within the field of psychometrics and ultimately deepen the understanding of undergraduate student engagement. By exploring the complexity and richness of this construct and using rigorous empirical methods, this research aims to empower educators and institutions with insights that can be applied to shape enriching and transformative learning experiences for diverse undergraduates.

In this empirical study, the multidimensionality underlying the construct of undergraduate student engagement was comprehensively explored. Using advanced statistical techniques such as full information item factor analysis (FIF), this study aims to utilize all available information from the observed data, including values that may be missing (Muthén & Muthén, 1998/2017). By using FIF, this research ensures that the exploration of the multidimensionality of student engagement is based on the most robust and accurate

technique, increasing the score validity and reliability of instruments. A transparent account of the steps taken to reveal the multidimensional nature of student engagement (Wong & Liem, 2022) is presented, with a particular focus on the utilization of FIF (Lee & Xu, 2003; Zhang et al., 2018) and between-item and within-item multidimensionality (Hartig et al., 2012; Hartig & Höhler, 2009).

This paper demonstrates that IRT provides a flexible model-based approach to examining the factor structure of instruments used in educational psychology research and offers an alternative approach to CFA for the dimensional assessment of psychological instruments. To finalize the reader's understanding of IRT and MIRT, in particular, CFA and MIRT methods are used to examine the dimensionality of the Student Engagement Scale, an instrument designed to operationalize student engagement, whether cognitive, behavioral, or social. The following section provides an overview of the main principles of IRT and the details of the unidimensional 2-parameter logistic (2-PL) IRT model for dichotomous data and Samejima's (1997) graded response model for categorical items (e.g., Likert scales). Additional aspects of IRT that align with the use of the MIRT model are reviewed, including item parameter estimation, ability estimation, and goodness of fit. Furthermore, an assessment of factor structure using the MIRT model is presented.

Item Response Theory

A participant's behavior in choosing a particular response to a problem or statement is a function that can be explained by two things: the character of the problem (statement) and the character of the participant. Participants' responses to the presented options can be scored dichotomously (0 = *false*, 1 = *true* or 0 = *no*, 1 = *yes*) or polytomously (0 = *strongly disagree*, 1 = *disagree*, 2 = *neither disagree nor agree*, 3 = *agree*, 4 = *strongly agree*). Modeling the interaction between the two characteristics is packaged in a nonlinear relationship on the same scale (θ) modeled in a function predicting the probability of giving a particular response.

IRT encompasses a comprehensive statistical framework that aims to capture the likelihood of an individual selecting a specific response to an item (Fan & Sun, 2013). In essence, IRT asserts that both item and individual attributes play a role in shaping item responses. Further, IRT multidimensionality allows for nuanced interpretations of test results, particularly in the social and behavioral sciences (Chung & Houts, 2020). Its application extends to predictive performance in test responses, where traits of individuals are quantified relative to their responses, demonstrating the crucial role of individual differences (Chang et al., 2022). Advanced algorithms like variational Bayesian inference in IRT highlight how large modern datasets can capture detailed behavioral nuances (Wu et al., 2020). Additionally, IRT models account for respondent behaviors like overreporting or underreporting, further emphasizing the interaction between individual attributes and item features (Leng et al., 2020). Comprehensive literature on IRT models underscores their broad applicability and depth of research in this field (Halpin, 2020).

The properties of the items, including item discrimination and threshold parameters, hold significant relevance. Item discrimination pertains to an item capability to differentiate among individuals along a spectrum of underlying traits (e.g., social engagement, cognitive engagement), such as distinguishing between students with varying levels of engagement. On the other hand, the item threshold signifies the juncture on the continuum of the latent trait at which an individual holds a .50 likelihood of choosing a specific response category. In instances involving dichotomously scored items, where responses are either true or false, the threshold gauges item difficulty, indicating the ease or difficulty with which respondents answered the item. Conversely, for psychological instruments frequently composed of ordered categorical items (e.g., Likert

scales), the threshold marks the point on the trait spectrum where an individual would have a .50 probability of opting for a particular response category. An individual's attribute manifests as his/her position on the measured trait (e.g., cognitive engagement, behavioral engagement, or social engagement), often referred to as proficiency, ability, or theta (denoted as θ).

IRT possesses several appealing characteristics for examining the psychometric attributes of psychological instruments (Bock & Gibbons, 2021; Hambleton, 2000; Reckase, 2009; Reise et al., 2018). Initially, IRT item parameter estimates are autonomous and unaffected by samples. Unlike the item discrimination and difficulty values derived from the classical test theory (Crocker & Algina, 1986), IRT item parameters remain uninfluenced by the underlying sample. Hence, an IRT model fit will produce consistent item parameter values, irrespective of the distribution of the sample (Embretson & Reise, 2000). This robust property of item parameters is a fundamental tenet of IRT. It underlies the computation of an individual's likelihood of responding correctly at a given level of ability or trait.

An individual's assessment of the measured trait (e.g., behavioral engagement, cognitive engagement) is unrelated to the specific items but any of item. Consequently, once a collection of items has been estimated (or calibrated), individual trait values can be ascertained by using the designated subset of items. Further, there is no necessity to administer an identical set of items to deduce individual trait estimates. In this scenario, a cluster of items can be chosen from the calibrated item pool and presented to individuals to ascertain their position on the quantified trait, forming the foundation of computerized adaptive testing.

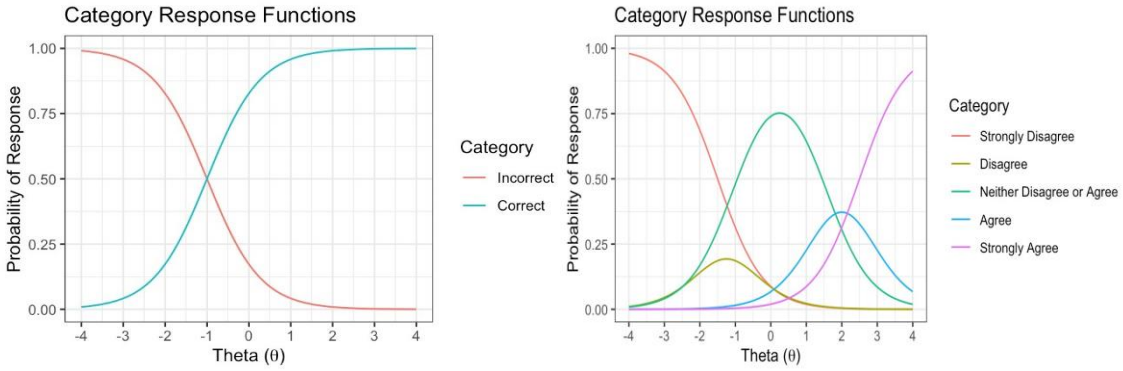
IRT calculates measurement errors for each attribute estimate. This stands in contrast to the standard error of measurement in classical test theory (CTT), which remains constant across the entire score distribution (Crocker & Algina, 1986). These constitute three evident advantages of employing IRT for item analysis and scoring. As with other statistical methodologies, the assumptions must undergo empirical evaluation to steer model selection determinations. These assumptions may only sometimes be relevant in all IRT applications. Hence, researchers should be familiar with the accessible data and models.

In this research the 2-parameter logistic (2PL) model (Hambleton et al., 1991) and the graded response model (GRM; Samejima, 1997) for didactic purposes are discussed. The 2PL item response theory model can be applied to dichotomous data, for example, 0 for *no* or *incorrect* and 1 for *yes* or *correct*. Meanwhile, the GRM can be applied to items scored polytomously, for example, 0 = *strongly disagree*, 1 = *disagree*, 2 = *neither disagree nor agree*, 3 = *agree*, and 4 = *strongly agree*. Understanding the simple unidimensional IRT model is a prerequisite for understanding the MIRT as an extended model (Reckase, 2009).

The 2PL model describes the probability of an individual choosing Response 1 as:

$$P(x = 1|\theta)\frac{1}{1 + e^{a_i(\theta - b_i)}}, \quad (1)$$

where $P(x = 1|\theta)$ represents the probability of choosing Response 1 (*correct* category or *yes* category) conditional on the individual's proficiency or latent trait position, a_i is the discrimination parameter, θ represents the proficiency trait, b_i is the item threshold, and e ~2.718. The subscript i indicates that each item has a unique discrimination and threshold parameter. Thus, the 2PL model is so named because it describes the probability of an examinee answering 1 based on two item parameters, that is, item discrimination and difficulty (b). Figure 1 (a) visualizes the probability of choosing the *correct* or *incorrect* response with a = .90 and b = –1.



(a) Two categories (b) Five categories

Figure 1
Item category response function

IRT models can also be applied to items with categorical response options or scored polytomously (van der Linden, 2016). Among these are the following: partial credit model (PCM; Masters, 1982), rating scale model (RSM; Andrich, 2010), generalized partial credit model (GPCM; Muraki, 1992), graded response model (GRM; Samejima, 1997), nominal response model (NRM; Bock, 1997b), nested logit model (NLM; Suh & Bolt, 2011), and multiple-categorical response model (MCRM; Thissen et al., 1989). Consulting Penfield's (2014) compact description of the polytomous item response model is suggested. A new, more complete, and comprehensive IRT model classification can be found in Kim et al. (2020). Collectively, these models are designed to predict the probability of a participant selecting a particular response category (e.g., *strongly disagree*, *disagree*, *agree*, *strongly agree*) for a statement in an item. For example, a student with high engagement with his/her college will have a high probability of responding *strongly agree* to the item "I am proud of my college." Samejima's GRM is an IRT model that can be applied to items scored polytomously. Specifically, it estimates the probability that individual *j* selects category *k* (e.g., *agree*, *strongly agree*) for item *i*:

$$P^*_{kji}\frac{1}{1+e^{-a_i(\theta-b_{ik})}}, \tag{2}$$

where P^*_{kji} is the probability of student *j* selecting category *k* or higher on item *i*, a_i is the item discrimination parameter (and is the same across all response categories), and b_{ik} is the threshold for reaching category *k*. For polytomous items, there are $k-1$ thresholds (one less than the number of response categories), meaning there are four thresholds (5-1) for a 5-point Likert scale. The four thresholds will cover the point on the scale where a student would choose a rating of *disagree* over *strongly disagree*, *neither disagree nor agree* over *disagree*, *agree* over *neither agree nor agree*, and *strongly agree* over *agree*. The probability of choosing a particular response category over choosing a lower response is:

$$P^*_{kni}=\frac{1}{1+e^{-a_i(\theta-b_{k-1})}}-\frac{1}{1+e^{-a_i(\theta-b_k)}}. \tag{3}$$

In this case, the term of the equation on the right of the minus sign is the probability of choosing the lower category (e.g., *disagree*), and the term on the left is the probability for the other category (e.g., *agree*). As

mentioned, the discrimination parameter a_i indicates that the model assumes that each category of items has the same discriminating power. Figure 1 (b) displays the category response function (CRF) for a polytomous item with four threshold parameters. Specifically, it depicts the probability of choosing one of the five possible response categories based on a given trait level. For this item, the discrimination parameter is .90, and the four threshold parameters are –1.5, –1, 1.5, and 2.5, respectively. As the model specifies, the categorical trace lines have the same slope and unique threshold parameters. The hypothetical trace lines in the figure can correspond to any item scored polytomously.

Several approaches were available for calibrating IRT item parameters. Expectation maximization (EM) or maximum marginal likelihood with expectation maximization (MML-EM; Bock, 1997b), quasi-Monte Carlo EM estimation (QMCEM; Jank, 2005), Markov chain Monte Carlo (MCMC; Edwards, 2010; Rabe-Hesketh et al., 2002), Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2010a), and stochastic EM (SEM; Nielsen, 2000; Zhang et al., 2018) can all be used as iterative procedures to estimate the item parameters. Chalmers (2012) suggested using the QMCEM when the dimensions are three or more. The type of estimation approach used to derive item parameter estimates will depend mainly on the statistical software package used for the IRT analysis. For example, the IRTPRO programs (Cai et al., 2018) implement the MML-EM procedure, while the "mirt" package in the R programming environment implements optional EM, QMCEM, MCEM, SEM, or MH-RM (Chalmers, 2012).

Various approaches were available to provide estimates of individual abilities or traits in IRT. These approaches fall under maximum likelihood or Bayesian procedures: the expected a posteriori (EAP), the maximum a posteriori (MAP), maximum likelihood (ML), weighted likelihood estimation (WLE), and the expected a posteriori for each sum score (EAPsum) (Warm, 1989). MLE seeks to determine the ability estimate, θ, that maximizes the likelihood of an individual's response pattern on a set of statement items. MLE minimum and maximum values can range from –∞ to +∞, so it is necessary to set a range of values, usually –3 to +3.

Bayesian methodologies encompass techniques such as MAP and EAP (Bock, 1997a). These approaches leverage a prior distribution of the ability distribution, drawing upon prior knowledge from the group of individuals whose scores are underestimated. Typically, scores are assumed to follow a normal distribution with a mean of 0 and a standard deviation of 1. Within Bayesian frameworks, a posterior distribution is generated for each individual, where the likelihood of the observed item response pattern is assessed at various ability levels (θ). EAP scores are derived from the posterior distribution mean, whereas MAP scores stem from the mode of this distribution, constrained within a score range of –3 to +3. Bayesian scores offer several appealing qualities over maximum likelihood estimation (such as reduced standard errors and the absence of extreme values), rendering them a more enticing choice.

In Bayesian statistics, estimating a test-taker's ability starts with a prior belief, updated by test responses to form a posterior distribution. This distribution merges the initial belief with the likelihood of the data. Maximum a posteriori (MAP) estimation pinpoints the ability level with the highest posterior probability. It involves adjusting a pretest belief about a student's ability based on their performance and then identifying the ability level where this revised belief peaks. Expectation a posteriori (EAP) estimation, however, accounts for the entire range of possible abilities. It calculates the expected ability value across the posterior distribution, effectively averaging all abilities, each weighted by probability. MAP concentrates on the most likely ability level, whereas EAP assesses the average of all potential levels, providing a more balanced estimate, particularly when the posterior distribution is skewed or ability estimates are uncertain. For foundational knowledge of Bayesian methods, Bergh et al. (2021) is recommended. Bock and Gibbons (2021) offer specific insights in the context of IRT. This summary delineates the differences between MAP and EAP in Bayesian inference within IRT, underscoring MAP focus on the most probable ability level and EAP comprehensive evaluation of all abilities.

Multidimensional Item Response Theory

The multidimensional item response theory (MIRT) expands the unidimensional IRT model by aiming to elucidate item responses based on an individual's positioning across multiple latent dimensions (Reckase, 2009). In practical research, the primary limitation of unidimensional models is their potential inability to accommodate widely used multidimensional measurement instruments adequately. As a result, advancements in MIRT and the growing availability of statistical software packages offer a valuable opportunity for applied researchers to develop an understanding of its application in assessing the psychometric performance of their scales. MIRT holds significant relevance, especially considering the intricate nature of the psychological constructs being investigated, such as how individuals approach learning, and the intricate interplay of personal and environmental factors influencing it. Analogous to other statistical modeling methods, MIRT contributes by enabling an exploration of the factors that contribute to an individual's response patterns to question items.

MIRT encompasses a wide-ranging set of probabilistic models developed to characterize an individual's likelihood of responding to an item, utilizing item parameters and latent traits. Specifically, MIRT positions an individual within a multidimensional space of latent traits hypothesized to influence item responses: $\theta_j = [\theta_{j1}, \theta_{j2}, \theta_{j3}, \ldots, \theta_{jM}]'$, where M denotes the count of unobserved latent dimensions required to model an individual's probable item responses. Two main categories of MIRT models are recognized: compensatory and noncompensatory (Reckase, 2009). Compensatory models allow a test-taker's higher positioning on one latent trait to potentially offset a lower positioning on another dimension, influencing the estimated probability of accurately responding to a question item. In contrast, noncompensatory (or partially compensatory) models restrict an individual's position across the multidimensional space to have no bearing on the likelihood of answering a question item. Within the existing literature, compensatory MIRT models are more frequently utilized. For dichotomous items, the probability of responding affirmatively to an item (e.g., *correct* or *yes*) can be formulated as:

$$P(u_i = 1|\theta_j, a_i, d_i) = \frac{e^{(a_i\theta_j + d_i)}}{1 + e^{(a_i\theta_j + d_i)}}, \tag{4}$$

where a_i represents the vector of item discrimination parameters (slope), indicating the probability of answering correctly associated with a change in a participant's position along m dimensions, and d_i corresponds to the item intercept parameter. Notably, the intercept d_i replaces the item threshold parameter (b) present in the unidimensional 2PL model and is not to be interpreted as a threshold (or difficulty level). Interested readers may refer to Reckase (2009, pp. 86-91) for a detailed presentation of the parameters of the multidimensional 2PL model. The exponent in Equation (4) can be expressed as:

$$a_i\theta_j{}' + d_i = a_{i1}\theta_{j1} + a_{i1}\theta_{j1} + \cdots + a_{im}\theta_{jm} + d_i. \tag{5}$$

In this context, a_{i1} denotes the slope or discrimination parameter for item i to a participant's position j on the first latent dimension, θ_1, and similarly for other up to m dimensions. Additionally, d_i represents the intercept. The mathematical expression of the MIRT model underscores its significance as a valuable psychometric instrument for estimating item and ability parameters across multiple dimensions (m dimensions).

The multidimensional graded response model (MGRM; Gibbons et al., 2007) can be written as:

$$P(y_{ij} = k|\theta_j) = \frac{1}{1 + e^{-\left(d_{j(k-1)} + a_i'\theta_j\right)}} - \frac{1}{1 + e^{-\left(d_{jk} + a_i'\theta_j\right)}}. \quad (6)$$

Here, *k* signifies the response category selected by individual *j* for item *i*. Like the unidimensional model, the ability estimates are not bounded and can span from negative to positive infinity, although they usually lie between –3 and +3. The MIRT model parameters are estimated using the same methodology as the unidimensional model, and checks for goodness of fit are grounded in the fit indices reported previously. The fundamental distinction between unidimensional IRT and MIRT models lies in the number of dimensions utilized to elucidate item responses.

Various models are available to describe the underlying factor structure of an assessment tool. As depicted in Figure 2, four potential factor structures for a 7-item measure are illustrated: a unidimensional model (Model A), a simple factor model (Model B), a correlated factor model (Model C) with two or three factors, a higher order factor model (Model D), a bifactor model (Model E), and a within-item model (Model F).

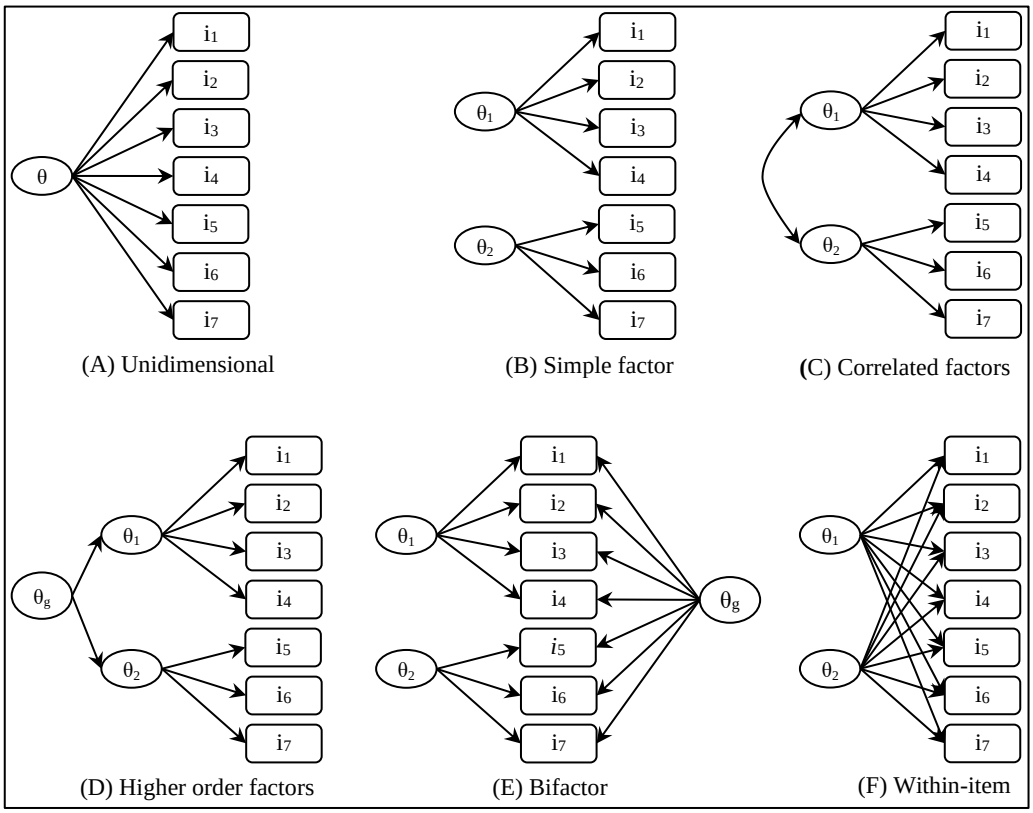


Figure 2
Alternative factor structures of a 7-item measure

The unidimensional, or single-factor, model is the most straightforward if all items gauge a single dimension. However, in many scenarios, the psychological concept of interest is theorized to be multidimensional. Consequently, capturing a multidimensional construct necessitates constructing scales that conceptualize items linked to multiple latent dimensions (e.g., cognitive and behavioral or behavioral and social). This theoretical framework prompts researchers to employ unrestricted methods like exploratory factor analysis or constrained

methods like confirmatory factor analysis to ascertain these subdomains' distinctiveness, serving modeling and score reporting purposes. Distinct subdomains might lead to reporting subscale scores, while strongly correlated factors may support utilizing a unified score.

In establishing the factorial validity of an assessment instrument, the choice of a model holds significant weight for researchers. Decisions regarding model selection should be guided by substantive theory and existing empirical evidence on the instrument's psychometric properties. Additionally, it is crucial to explore various factor structures to rule out alternative interpretations of the instrument underlying structure. This could encompass unidimensional, correlated factor, bifactor, and within-item models.

Researchers need to recognize and comprehend the commonalities among different statistical models. Disregarding a more appropriate model during the selection process could lead to erroneous conclusions (Lee et al., 2019). For instance, correlated 2- or 3-factor models (as exemplified by Model C, Figure 2) are founded on the premise that items measure distinct yet interconnected latent dimensions. In cases where factor correlations approach unity, a 1-factor (unidimensional) model might offer a better fit for the data, potentially challenging the concept of different latent dimensions. Ultimately, researchers should consider the intricacies of the models, grounded in theory and empirical evidence, to choose the one that best aligns with the nature of the investigated instrument.

In contrast, a higher-order model may be more appropriate if a factor hierarchy can explain the relationships between factors (see Model D, Figure 2). In recent years, in education (Gibbons & Hedeker, 1992) and psychology (Gibbons et al., 2007) literature, the bifactor structure has gained increasing attention (see Model D, Figure 2; Reise, 2012). A bifactor structure states that the interrelationships among all items are explained by a primary dimension with a conceptually grouped subset of items related to a particular subdomain. Gibbons et al. (2007) developed a full-item information bifactor model for multilevel response data. The fundamental assumptions of this model are that all items correspond to one central dimension and one subdomain and that the dimensions are uncorrelated or orthogonal.

Further, Cai (2010b) proposed a 2-level model that extended the 2-factor model and demonstrated its application for modeling complex and longitudinal data structures. Despite the apparent differences between the models, the literature shows similarities. Rindskopf (2012) showed the relationship between unidimensional models, correlated factors, higher order models, and bifactor models. In IRT literature, Rijmen (2010) showed how second order models are equivalent to testlet models, and both are restricted bifactor models. Thus, through this comprehensive examination, this research aims to clarify the intricate relationships and applicabilities of various factor models in MIRT to student engagement measurement modeling, guided by empirical evidence and theoretical underpinnings.

Methods

Participants

The dataset employed in the study consisted of responses from university students and aimed to evaluate student engagement in higher education. A total of 657 undergraduate students responded to the questionnaire; 427 were female, and the rest were male. Their ages ranged from 18 to 23 years (M = 20, SD = 1.72). The participants were from 13 universities under the Ministry of Religious Affairs (MORA) located in the Indonesian cities of Aceh, Bandung, Banjarmasin, Makassar, Malang, Medan, Palembang, Pekanbaru, Palangkaraya,

Pamekasan, Purwokerto, Salatiga, and Surabaya. Informed consent was obtained by informing the students that they were giving consent by completing the online survey and that their participation was voluntary.

Measures

The Student Engagement Scale is a 20-item measurement tool for assessing undergraduate student engagement. The scale was developed based on a conceptual review of work engagement, with adjustments made to fit the context of undergraduate students in higher education. Given the absence of empirical evidence regarding the scale's internal structure, this instrument emerged as a suitable candidate for dimensional assessment. This scale is an extension of Ridho's (2023) work, implemented in the context of students' performance in completing their undergraduate studies as "student engagement." The scale includes five subscales: cognitive engagement, emotional engagement, physical engagement, behavioral engagement, and social engagement. Each item in the five subscales asks for five graded responses (0 = *strongly disagree*, 1 = *disagree*, 2 = *neither disagree nor agree*, 3 = *agree*, 4 = *strongly agree*).

Data Analysis

The Multidimensional item response theory (MIRT; Reckase, 2009) was used to evaluate the factor structure of the instruments in this study. As a first step, descriptive statistics were used for data screening purposes. An initial CTT-based item discrimination analysis (r_{ir}) was conducted to inform the subsequent IRT-based item parameter calibration. Additionally, exploratory factor analysis and principal component analysis were employed to validate the multidimensional nature of the engagement construct, following established practices (Fan & Sun, 2013; Hambleton et al., 1991). To corroborate these results, in parallel it was tested whether a 1-, 2-, or 3-factor model based on a polytomous graded response model (GRM; Gibbons et al., 2007; Samejima, 1997), corresponding to the form of response offered (*strongly disagree* to *strongly agree*), as appropriate. This research tested the proposed models in the next phase to best fit the data.

Five models were confirmed in this study. Model 1 is a multidimensional simple 3-factor, Model 2 is a multidimensional correlated 3-factor, Model 3 is a multidimensional 3-bifactor (one primary factor and two specific factors), Model 4 is a multidimensional bifactor with one primary factor and three specific factors, and Model 5 is a within-item 3-factor. Each model forms the basis for evaluating the extent to which the factor structure of the instrument is unidimensional, composed of distinct cognitive, behavioral, and social engagement dimensions, or complex with items related to the primary dimension and domain-specific cognitive, behavioral, or social factors, or complex with items related to the whole dimensions. The "mirt" package, Version 1.39 (Chalmers, 2012) in R environment (R Core Team, 2022), was used to analyze the data modeling from students' responses to the questionnaire.

For robust parameter estimation, the QMCEM estimation method was employed within the "mirt" package. The model fit was evaluated using various indices: Akaike information criteria (AIC; Akaike, 1998), Bayesian information criteria (BIC; Schwarz, 1978), M2 (Maydeu-Olivares, 2013), root-mean-square error of approximation (RMSEA; Steiger, 1990, 2016; Yin et al., 2023), standardized root-mean-square residual (SRMR; Pavlov et al., 2021), Tucker-Lewis index (TLI; Cai et al., 2023; Tucker & Lewis, 1973), and comparative fit index (CFI; Bentler, 1990; Hu & Bentler, 1999). Consistent with Hu and Bentler's (1999) recommendations, an acceptable model fit was indicated by SRMR $\leq$ .08, RMSEA $\leq$ .05, TLI $\geq$ .90, and

CFI ≥ .95. Furthermore, smaller AIC and BIC values, indicative of a better fit, were considered (Huang, 2017). Additionally, the differences in CFI between a model and its augmented version, reflecting an increase in parameters, were also taken into account, as suggested by Cheung and Rensvold (2002).

RESULTS

The item-rest correlation, r_{ir}, of Item i12 illustrates a deviation given that it has a value of –.01, indicating a meager, even harmful, discriminating power, which also means that the item is not at all in line with the other items in the scale. Therefore, the subsequent analysis did not include Item i12. Next, the frequency distribution reported that the item responses had a negatively skewed distribution. Average skewness is –.61 with a range of –1.69 (Item i08) to –.07 (Item i15) (see Table 1).

Table 1
Item descriptive statistics

Item	p1	p2	p3	p4	p5	*M*	*SD*	Min	Max	Skew	Kurt	r_{ir}
i01	.03	.09	.21	.40	.28	3.81	1.02	1	5	–0.71	–0.01	.60
i02	.02	.07	.18	.45	.28	3.91	0.95	1	5	–0.80	0.34	.70
i03	.01	.06	.22	.48	.23	3.86	0.88	1	5	–0.70	0.46	.66
i04	.03	.05	.15	.37	.39	4.06	1.00	1	5	–1.06	0.77	.37
i05	.03	.06	.30	.40	.21	3.70	0.96	1	5	–0.54	0.11	.71
i06	.03	.08	.29	.39	.21	3.66	1.00	1	5	–0.53	–0.07	.73
i07	.03	.07	.27	.42	.22	3.73	0.97	1	5	–0.64	0.20	.70
i08	.01	.01	.09	.29	.59	4.43	0.82	1	5	–1.69	3.26	.35
i09	.03	.07	.36	.37	.17	3.58	0.95	1	5	–0.39	0.02	.59
i10	.04	.06	.25	.36	.29	3.79	1.07	1	5	–0.75	0.10	.56
i11	.02	.07	.39	.39	.13	3.55	0.87	1	5	–0.29	0.13	.65
i13	.02	.06	.34	.40	.18	3.65	0.92	1	5	–0.39	0.03	.49
i14	.01	.07	.31	.45	.16	3.68	0.87	1	5	–0.42	0.07	.60
i15	.04	.21	.46	.25	.04	3.04	0.89	1	5	–0.07	–0.14	.50
i16	.03	.14	.50	.27	.06	3.19	0.84	1	5	–0.08	0.22	.50
i17	.02	.05	.23	.48	.22	3.82	0.90	1	5	–0.74	0.64	.39
i18	.03	.09	.33	.42	.12	3.52	0.93	1	5	–0.49	0.14	.53
i19	.00	.03	.20	.54	.22	3.94	0.77	1	5	–0.56	0.53	.57
i20	.03	.06	.28	.46	.18	3.70	0.92	1	5	–0.66	0.48	.58

Note. p1 = *strongly disagree* proportion response; p2 = *disagree* proportion response; p3 = *neither disagree nor agree* proportion response; p4 = *agree* proportion response; p5 = *strongly agree* proportion response; Skew = Skewness; Kurt = Kutosis; r_{ir} = item-rest correlation.

Table 1 reports descriptive statistics for the scale items. The average response was *neither disagree nor agree* to *agree* across the items. An examination of the minimum and maximum values indicates that there is a range of restrictions for each item. The item-rest correlation, r_{ir}, ranged from .35 (Item i08) to .73 (Item i06) with a mean of .57 (*SD* = 0.11). Cronbach's alpha coefficients for the scale total scores were .91, .90 for the cognitive engagement subscale, .81 for the behavior engagement subscale, and .75 for the social engagement

subscale. The item analysis results and Cronbach's alpha coefficients provided relevant information to guide a model-based approach to assessing scale factor structure (e.g., relationships among item sets).

Exploratory Model

Exploration was conducted on the response data. As presented in Table 2 and in the Appendix (Table 1A), the goodness of fit suggests that the data fit the 3-factor model better than the 2- and 1-factor models. It is supported by the model fit indices (AIC, BIC, M2, RMSEA, SRMR, TLI, and CFI). The 3-factor model produces the smallest AIC, BIC, and M2 indices, RMSEA and SRMR are less than .05, and TLI and CFI are more than .90, proving that the internal structure of the engagement instrument is 3-factor multidimensional. The significant likelihood ratio test with $p < .01$ in Table 3 (see also Table 2A in the Appendix) also corroborates the evidence of the precision of the 3-factor model over its 2- and 1-factor counterparts. Additionally, the CFI difference between the 2-factor and 3-factor ΔCFI > .01 indicates that the 2-factor model does not explain the data better than the 3-factor model. Therefore, the following confirmation process involved some variation of the 3-factor structural model.

Table 2
Fit comparison of confirmatory factor IRT models

Model	AIC	BIC	M2	*df*	RMSEA	SRMR	TLI	CFI
Model 1	27466.98	27893.31	385.262	95	.07	.26	.81	.84
Model 2	26868.87	27308.66	307.749	92	.06	.28	.85	.88
Model 3	26671.86	27183.46	171.385	76	.04	.04	.92	.95
Model 4	26647.61	27159.21	183.123	76	.05	.05	.91	.94
Model 5	26634.93	27231.79	121.055	57	.04	.04	.93	.96

Note. AIC = Akaike information criteria; BIC = Bayesian information criteria; M2 = M2 statistic; *df* = degrees of freedom; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; TLI = Tucker-Lewis index; CFI = comparative fit index.

Table 3
Likelihood ratio test model comparison

Model	*N* Par	logLik	χ^2	*df*	*p*
Model 1	95	–13638.49	–	–	–
Model 2	98	–13336.43	604.12	3	0
Model 3	114	–13221.93	229.01	16	0
Model 4	114	–13209.81	24.25	0	–
Model 5	133	–13184.47	50.68	19	0

Note. *N* Par = number of parameters estimated; logLik = loglikelihood.

Table 3A in the Appendix reports the factor loadings of each item of the student engagement instrument across the three models explored, suppressing loadings < .30. In the 1-factor model, factor loadings ranged from .39 (Item i08) to .90 (Item i06) with mean .66 and standard deviation 0.16. Some items (e.g.,

i08, i09, and i10) have sufficient factor loadings on two dimensions in the 2- and 3-factor models. This information shows variance sharing in items with more than one factor load.

Confirmatory Model

As explained in the Methods section, five multidimensional models were compared in this study. Model 1 is a simple 3-factor model, Model 2 is a correlated 3-factor model, Model 3 is a bifactor model with one primary dimension and two specific dimensions, Model 4 is a bifactor model with one primary dimension and three specific dimensions, and Model 5 is a within-item multidimensional model where three dimensions account for all items composing the student engagement scale. The data fit presented in Table 2 shows that Model 3, Model 4, and Model 5 fulfill the absolute data fit criteria (SRMR < .08; RMSEA < .05; TLI > .90). All data fit indices (except BIC) informed that Model 5 was the best-fit model for the data. The likelihood ratio test in Table 3 also shows that Model 5 fits the data best compared to the other models being compared.

Table 4 reports the factor loadings for each cognitive (F1), emotional (F2), and social (F3) dimension. The discriminating power parameters corresponding to these dimensions are the discriminating parameters of the cognitive (a_1), emotional (a_2), and social (a_3) dimensions. The factor loadings correlated with the discriminating power because they can be converted to each other. The higher the factor load, the higher the discriminating power. The following parameter, location or easiness (d), is the ease of each response category. For example, d_1 is the location between *strongly disagree* and *disagree*, whereas d_4 parameterizes the location between *agree* and *strongly agree*. The location of each item will generally get smaller ($d_4 < d_3 < d_2 < d_1$). Alternatively, these parameters can be converted into a multidimensional difficulty parameter (MDIFF). The triple discriminating power for each item can be converted into a multidimensional discriminating parameter (MDISC), as reported in Table 4A in the Appendix.

Table 4
Standardized within-item factor loading, item discrimination, and item easiness on three dimensions

Item	F1	F2	F3	a1	a2	a3	d1	d2	d3	d4
i01	.48	.57	.15	1.26	1.50	0.39	5.27	3.14	1.18	–1.48
i02	.56	.64	.17	1.92	2.20	0.59	7.61	4.68	2.05	–1.85
i03	.54	.63	.14	1.66	1.95	0.43	7.24	4.54	1.64	–2.23
i04	.26	.37	.12	0.50	0.71	0.23	3.98	2.74	1.35	–0.50
i05	.70	.58	.17	3.00	2.49	0.73	8.67	5.71	1.14	–3.33
i06	.69	.60	.19	3.21	2.80	0.89	8.99	5.62	1.10	–3.55
i07	.62	.58	.21	2.26	2.12	0.77	6.98	4.62	1.21	–2.70
i08	.06	.34	.28	0.11	0.64	0.54	4.60	3.86	2.23	0.41
i09	.28	.54	.31	0.65	1.27	0.73	4.62	2.94	0.20	–2.19
i10	.25	.50	.31	0.55	1.11	0.69	3.90	2.78	0.80	–1.19
i11	.16	.68	.40	0.46	1.97	1.16	6.34	3.90	0.16	–3.23
i13	.03	.42	.53	0.08	0.96	1.23	5.10	3.22	0.45	–2.07
i14	.01	.63	.51	0.02	1.84	1.49	6.98	4.05	0.78	–2.86
i15	.12	.33	.56	0.28	0.75	1.28	4.15	1.49	–1.22	–4.03
i16	.03	.46	.51	0.03	1.13	1.21	4.87	2.28	–1.03	–3.87

(table 4 continues)

Table 4 (continued)

Item	F1	F2	F3	a1	a2	a3	d1	d2	d3	d4
i17	.14	.14	.54	0.28	0.29	1.12	4.55	3.05	1.06	–1.61
i18	.36	.08	.74	1.07	0.23	2.24	5.97	3.59	0.38	–3.45
i19	.24	.23	.70	0.64	0.63	1.89	7.68	4.92	1.83	–2.06
i20	.44	.14	.65	1.24	0.41	1.85	5.71	3.98	0.98	–2.55

Note. F1 = loading factor for the cognitive dimension; F2 = loading factor for the emotional dimension; F3 = loading factor for the social dimension; a1 = discrimination parameter for the cognitive dimension; a2 = discrimination parameter for the emotional dimension; a3 = discrimination parameter for the social dimension; d1 = easiness parameter for choosing Category 2 (*disagree*); d2 = easiness parameter for choosing Category 3 (*neither disagree nor agree*); d3 = easiness parameter for choosing Category 4 (*agree*); d4 = easiness parameter for choosing Category 5 (*strongly agree*).

Figure 3 shows the item vector and participant parameters in the same 3-dimensional space scale. Figure 3(d) describes all items vector, Figure 3(a) for the cognitive items vector, Figure 3(b) for the behavioral items vector, and Figure 3(c) depicts the social items vector. Meanwhile, Figure 3(e) shows the 657 participants' positions in a 3-dimensional space, indicating the relative position of each when projected on the axes of the cognitive, behavioral, and social dimensions.

Discussion

The present study delved into applying the multidimensional item response theory (MIRT) to scrutinize the internal structure of a construct, specifically in the context of exploring and confirming dimensions of student engagement. The outcomes of the analysis provided valuable insights into the nature of this construct, shedding light on its multidimensional facets. This section discusses the key findings, offers visual representations of the multidimensional structure, and underscores the significance of MIRT in psychometric assessment.

The study primary objective was to unravel the latent dimensions inherent in student engagement. Drawing on the multidimensional approach validated by C.-Y. Chen et al. (2018) in their examination of the ages and stages questionnaires and the insights of Carlucci et al. (2023) on the state-trait inventory for cognitive and somatic anxiety, the results exhibited robust support for the multidimensional structure of the scale, revealing that the items tapped into multiple dimensions within the construct — this multidimensionality, akin to the findings in both C.-Y. Chen et al. (2018) and Carlucci et al. (2023) further validated confirmatory full-information item factor analysis outcomes.

The exploratory full-information item factor analysis comparing one, two, or three dimensions reveals that student engagement is not a monolithic construct but a multifaceted one comprising three interrelated dimensions. These dimensions, reflecting the interconnected nature of child development and anxiety constructs found in the earlier studies, could encompass cognitive, emotional, and behavioral aspects. This highlights how students interact with the learning process and their educational environment, similar to the multidimensional interactions observed in child development and anxiety assessments.

This multidimensional perspective of student engagement, mirroring the complex structures unraveled in the ASQ-3 and STICSA scales, has implications for educators and policymakers. It underscores the importance of addressing the various facets of engagement to promote holistic student development and improved learning outcomes, much like the implications drawn from C.-Y. Chen et al. (2018) and Carlucci et al. (2023) for their respective fields.

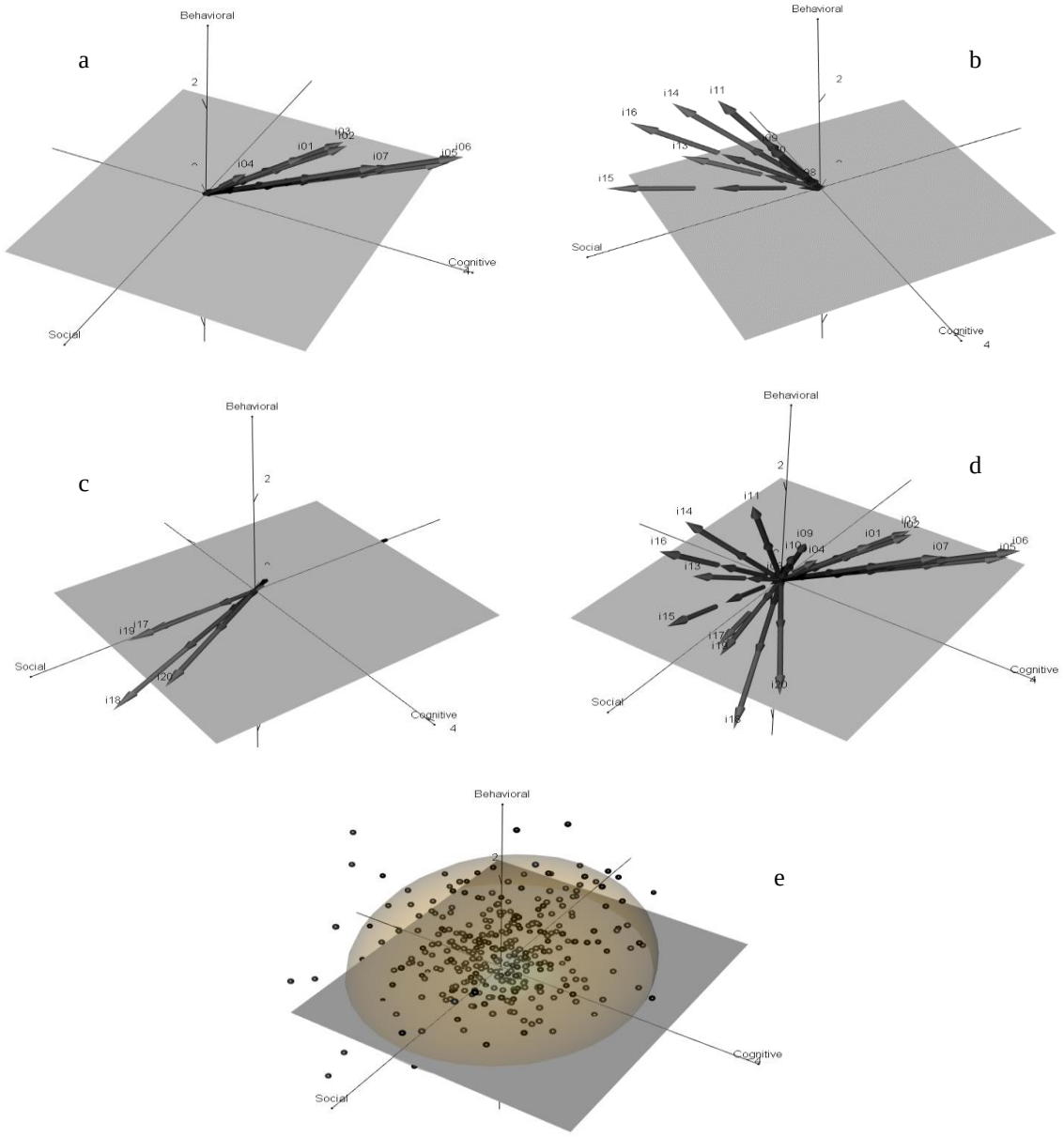


Figure 3
Item and participants' parameters in 3-dimensional space

The application of exploratory multidimensional item factor analysis lends empirical support to the contention that student engagement is inherently multidimensional. By scrutinizing the responses to various engagement-related items, researchers gained insights into the underlying structure of this construct, akin to the approach used in the studies by C.-Y. Chen et al. (2018) and Carlucci et al. (2023). These studies emphasized the need for a nuanced understanding of complex constructs, a principle echoed in our findings. The emergence of distinct factors within student engagement underscores the idea that it cannot be reduced to a singular, uniform concept. Instead, it manifests as a complex interplay of several dimensions, each contributing

uniquely to the overall engagement experience, mirroring the multidimensionality uncovered in child development and anxiety assessments.

Referring to the results in Table 3A in the Appendix, there was evidence that some items reveal one dimension, while others contain 2-dimensional loadings. Item 1 ("I feel energized to complete my study"), Item 2 ("I really pay attention to my study"), Item 3 ("I devote all my intellectual effort to my study"), Item 5 ("I am deeply excited with my study"), Item 6 ("I am enthusiastic about my study"), and Item 7 ("I find my study meaningful") contain cognitive and behavioral dimensions. Item 8 ("An unfinished assignment makes me always think about it") contains only one behavioral dimension. Meanwhile, Item 9 ("I am voluntarily increasing my study time"), Item 10 ("Even though I am tired, I sincerely attend when there are additional hours of lectures"), Item 11 ("I am courageous in lectures"), Item 13 ("When I do not understand, I ask questions in class discussions"), Item 14 ("I often do extra initiatives to complete coursework"), Item 15 ("I always actively take a role for my class"), and Item 16 ("I do more than expected by the lecturer") reveal both behavioral and social dimensions. Next, Item 17 ("I am used to cooperating with other students when I have problems") and Item 19 ("I develop a good relationship with my lecturers") each only reveal the social dimension, while Item 18 ("I belong to this campus") and Item 20 ("Being a student in this campus makes me feel passionate") contain both cognitive and social dimensions. This indicates a complex and interconnected structure, similar to the multidimensional constructs found in developmental and anxiety scales. Additionally, certain items exclusively revealing the social dimension and others containing both cognitive and social dimensions further validate the multidimensionality of student engagement, akin to the findings in the studies above.

The findings from this study, in conjunction with those of C.-Y. Chen et al. (2018) and Carlucci et al. (2023), reinforce the intricate nature of psychological constructs, particularly in the context of student engagement. The multidimensional structure identified in this study echoes the complexities observed in child development and anxiety assessments, underscoring a broader trend in psychological research that recognizes the multifaceted nature of human behavior and cognition. This trend moves away from oversimplified, unidimensional interpretations, advocating for a more nuanced understanding of psychological phenomena. The presence of interrelated cognitive, emotional, and behavioral dimensions in student engagement parallels findings in other domains, suggesting a universal need for multifaceted analytical approaches in psychological assessments.

The paper provision of vector depictions in a 3-dimensional space added a visual dimension to the results, aiding readers in grasping the multidimensional nature of the construct. Such visualizations offer an intuitive representation of how the items are situated within the broader cognitive, behavioral, and social dimensions. This approach reflects the interconnected nature of these dimensions, as seen in child development and anxiety assessments, and enhances the accessibility of the findings, making them more comprehensible to a broader audience. The emergence of distinct yet interrelated dimensions in student engagement challenges the notion of it being a singular, uniform concept. It underscores the complexities observed in similar psychological constructs and suggests a universal need for multifaceted analytical approaches in psychological assessments.

Conclusion

In conclusion, this study has underscored the utility of the multidimensional item response theory (MIRT) in both exploratory and confirmatory analyses of constructs like student engagement. Identifying and validating multidimensional factors within the construct contributes to a deeper understanding of its complexity.

The within-item multidimensional structure and the visual depictions further enhance the richness of the findings. As the study advocates, MIRT serves as an invaluable tool in the arsenal of psychometric assessments, enabling researchers and practitioners to unravel the multidimensional fabric of various constructs.

The results of this study have practical implications for education. Addressing cognitive, behavioral, and social engagement is critical to fostering a holistic learning environment, and educators are encouraged to use integrated strategies, such as challenging curriculum for cognitive engagement, participatory teaching for behavioral engagement, and collaborative activities for social engagement. This ensures the activation and nurturing of all aspects of student engagement. In addition, in recognizing the varying levels of engagement across different dimensions, educators should adopt different teaching techniques tailored to the engagement profiles of diverse student groups. This pedagogical approach ensures that each student's unique engagement needs are addressed. By embracing these multidimensional strategies, educators can create a more inclusive and effective learning environment that accommodates the preferences and needs of diverse students.

Acknowledgments

Thanks to all the participants of this study for their time and willingness to share their experiences. Their contributions were invaluable in helping me understand this topic and draw meaningful conclusions.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Akaike, H. (1998). Factor analysis and AIC. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 371-386). Springer. https://doi.org/10.1007/978-1-4612-1694-0_29

Andrich, D. (2010). Sufficiency and conditional estimation of Person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292-308. https://doi.org/10.1007/s11336-010-9154-8

Ben-Eliyahu, A., Moore, D., Dorph, R., & Schunn, C. D. (2018). Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. *Contemporary Educational Psychology*, *53*, 87-105. https://doi.org/10.1016/j.cedpsych.2018.01.002

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bergh, D. v. d., Clyde, M. A., Gupta, A. R. K. N., de Jong, T., Gronau, Q. F., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2021). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*, *53*(6), 2351-2371. https://doi.org/10.3758/s13428-021-01552-2

Bock, R. D. (1997a). A brief history of item theory response. *Educational Measurement: Issues and Practice*, *16*(4), 21-33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Bock, R. D. (1997b). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer.

Bock, R. D., & Gibbons, R. D. (2021). *Item response theory*. John Wiley & Sons.

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International Journal of Educational Technology in Higher Education*, *17*(2), 1-30. https://doi.org/10.1186/s41239-019-0176-8

Cai, L. (2010a). Metropolis-Hastings Robbins-Monro Algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307-335. https://doi.org/10.3102/1076998609353115

Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581-612. https://doi.org/10.1007/s11336-010-9178-0

Cai, L., Chung, S. W., & Lee, T. (2023). Incremental model fit assessment in the case of categorical data: Tucker-Lewis index for item response theory modeling. *Prev Sci*, *24*(3), 455-466. https://doi.org/10.1007/s11121-021-01253-4

Cai, L., Thissen, D., & du Toit, S. (2018). *IRTPRO (item response theory for patient-reported outcomes)* [Computer software]. Scientific Software International, Inc.

Carlucci, L., Innamorati, M., Ree, M., D'Ignazio, G., & Balsamo, M. (2023). Measuring state and trait anxiety: An application of multidimensional item response theory. *Behavioral Sciences*, *13*(8), Article 628. https://doi.org/10.3390/bs13080628

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chang, J. C., Porcino, J., Rasch, E. K., & Tang, L. (2022). Regularized Bayesian calibration and scoring of the WD-FAB IRT model improves predictive performance over marginal maximum likelihood. *PLoS One*, *17*(4), Article e0266350. https://doi.org/10.1371/journal.pone.0266350

Chen, C.-Y., Xie, H., Clifford, J., Chen, C.-I., & Squires, J. (2018). Examining internal structures of a developmental measure using multidimensional item response theory. *Journal of Early Intervention*, *40*(4), 1-17. https://doi.org/10.1177/1053815118788063

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, *83*(3), 538-562. https://doi.org/10.1007/s11336-018-9610-4

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Childs, R. A., & Oppler, S. H. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high-stakes testing program. *Educational and Psychological Measurement*, *60*(6), 939-955. https://doi.org/10.1177/00131640021971005

Chung, S., & Houts, C. (2020). FlexMIRT: A flexible modeling package for multidimensional item response models. *Measurement: Interdisciplinary Research and Perspectives*, *18*(1), 40-54. https://doi.org/10.1080/15366367.2019.1693825

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston Inc.

Edwards, M. C. (2010). A Markov Chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474-497. https://doi.org/10.1007/s11336-010-9161-9

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341-349. https://doi.org/10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. Lawrence Erlbaum Associates Inc.

Fan, X., & Sun, S. (2013). Item response theory. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 45-67). SensePublishers. https://doi.org/10.1007/978-94-6209-404-8_3

Fredricks, J. A., Hofkens, T. L., & Wang, M.-T. (2019). Addressing the challenge of measuring student engagement. In K. A. Renninger & S. E. Hidi (Eds.), *The Cambridge handbook of motivation and learning* (pp. 689-712). Cambridge University Press. https://doi.org/10.1017/9781316823279.029

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Ellen Frank, V. G., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4-19. https://doi.org/10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436. https://doi.org/10.1007/BF02295430

Groccia, J. E. (2018). What is student engagement? *New Directions for Teaching and Learning*, *154*, 11-20. https://doi.org/10.1002/tl.20287

Halpin, P. F. (2020). A review of handbook of item response theory (Vol. 1). *Journal of Educational and Behavioral Statistics*, *46*(4), 519-522. https://doi.org/10.3102/1076998620978551

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, *38*(9), II60-II65. https://doi.org/10.1097/00005650-200009002-00009

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publication Inc.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*(4), 665-686. https://doi.org/10.1177/0013164411430707

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*, 57-63. https://doi.org/10.1016/j.stueduc.2009.10.002

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

Huang, P.-H. (2017). Asymptotics of AIC, BIC, and RMSEA for model selection in structural equation modeling. *Psychometrika*, *82*(2), 407-426. https://doi.org/10.1007/s11336-017-9572-y

Inman, R. A., Moreira, P. A. S., Cunha, D., & Castro, J. (2020). Assessing the dimensionality of the student school engagement survey: Support for a multidimensional bifactor model. *Revista de Psicodidáctica (English ed.)*, *25*(2), 109-118. https://doi.org/10.1016/j.psicoe.2020.03.001

Jank, W. (2005). Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Computational Statistics & Data Analysis*, *48*(4), 685-701. https://doi.org/https://doi.org/10.1016/j.csda.2004.03.019

Kim, S.-H., Kwak, M., Bian, M., Feldberg, Z., Henry, T., Lee, J., Ölmez, İ. B., Shen, Y., Tan, Y., Tanaka, V., Wang, J., Xu, J., & Cohen, A. S. (2020). Item response models in psychometrika and psychometric textbooks. *Frontiers in Education*, *5*, Article 63. https://www.frontiersin.org/articles/10.3389/feduc.2020.00063

Kruglova, N., Dykhovychnyi, O., & Lysenko, D. (2021). Application of IRT and MIRT models to analysis of analytical geometry tests. *Adaptive Systems of Automatic Control*, *1*(38), 36-49. https://doi.org/10.20535/1560-8956.38.2021.233179

Lee, H. R., Lee, S., & Sung, J. (2019). The impact of ignoring multilevel data structure on the estimation of dichotomous item response theory models. *International Journal of Assessment Tools in Education*, *6*(1), 92-108. https://doi.org/10.21449/ijate.523586

Lee, S.-Y., & Xu, L. (2003). On local influence analysis of full information item factor models. *Psychometrika*, *68*(3), 339-360. https://doi.org/10.1007/bf02294731

Leng, C.-H., Huang, H.-Y., & Yao, G. (2020). A social desirability item response theory model: Retrieve-deceive-transfer. *Psychometrika*, *85*(1), 56-74. https://doi.org/10.1007/s11336-019-09689-y

Li, J., & Xue, E. (2023). Dynamic interaction between student learning behaviour and learning environment: Meta-analysis of student engagement and its influencing factors. *Behavioral Sciences (Basel)*, *13*(1). Article 59. https://doi.org/10.3390/bs13010059

Liu, Y., Magnus, B., O'Connor, H., & Thissen, D. (2018). Multidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The wiley handbook of psychometric testing* (pp. 445-493). John Wiley & Sons Ltd. https://doi.org/10.1002/9781118489772.ch16

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71-101. https://doi.org/10.1080/15366367.2013.831680

Mikkonen, K., Tomietto, M., & Watson, R. (2022). Instrument development and psychometric testing in nursing education research. *Nurse Education Today*, *119*, Article 105603. https://doi.org/10.1016/j.nedt.2022.105603

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. https://doi.org/10.1177/014662169201600206

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8 ed.). Muthén & Muthén. [Original work published 1998].

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, *6*(3), 457-489. https://doi.org/10.2307/3318671

Nkomo, L. M., Daniel, B. K., & Butson, R. J. (2021). Synthesis of student engagement with digital technologies: a systematic review of the literature. *International Journal of Educational Technology in Higher Education*, *18*, 1-26. https://doi.org/10.1186/s41239-021-00270-1

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, *183*, Article 111114. https://doi.org/10.1016/j.paid.2021.111114

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36-48. https://doi.org/10.1111/emip.12023

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1-21. https://doi.org/10.1177/1536867X0200200101

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor modeling and the evaluation of scale scores. In *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vols. 1-2, pp. 677-707). Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch22

Ridho, A. (2023). Work engagement: A conceptual review [Keterikatan Kerja: Sebuah Reviu Konseptual]. *Buletin Psikologi*, *31*(1), 56-79. https://doi.org/10.22146/buletinpsikologi.55589

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361-372. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A., Curby, T. W., & Abry, T. (2015). To what extent do teacher-student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning? *Journal of Educational Psychology*, *107*(1), 170. https://doi.org/10.1037/a0037252

Rindskopf, D. (2012). Next steps in Bayesian structural equation models: Comments on, variations of, and extensions to Muthén and Asparouhov (2012). *Psychological Methods*, *17*(3), 336-339. https://doi.org/10.1037/a0027130

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108-116. https://doi.org/10.7334/psicothema2013.260

Salmela-Aro, K., Tang, X., Symonds, J., & Upadyaya, K. (2021). Student engagement in adolescence: A scoping review of longitudinal studies 2010-2020. *Journal of Research on Adolescence*, *31*(2), 256-272. https://doi.org/10.1111/jora.12619

Samejima, F. (1997). Graded response model. In W. J. Van Linden & R. K. Hambleton (Eds.), *Handbook of modern item response* (pp. 85-100). Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464. https://doi.org/10.2307/2958889

Siu, O. L., Lo, B. C. Y., Ng, T. K., & Wang, H. (2023). Social support and student outcomes: The mediating roles of psychological capital, study engagement, and problem-focused coping. *Current Psychology*, *42*(4), 2670-2679. https://doi.org/10.1007/s12144-021-01621-x

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173-180. https://doi.org/10.1207/s15327906mbr2502_4

Steiger, J. H. (2016). Notes on the Steiger-Lind (1980) handout. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 777-781. https://doi.org/10.1080/10705511.2016.1217487

Suh, Y., & Bolt, D. M. (2011). A Nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*(2), 188-205. https://doi.org/10.1111/j.1745-3984.2011.00139.x

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247-260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442-1455. https://doi.org/10.1037/pas0000597

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10. https://doi.org/10.1007/BF02291170

van der Linden, W. J. (2016). Introduction. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 1-10). Taylor & Francis Group, LLC.

Wang, M.-T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The math and science engagement scales: Scale development, validation, and psychometric properties. *Learning and Instruction*, *43*, 16-26. https://doi.org/10.1016/j.learninstruc.2016.01.008

Wang, M.-T., & Hofkens, T. L. (2020). Beyond classroom academics: A school-wide and multi-contextual perspective on student engagement in school. *Adolescent Research Review*, *5*(4), 419-433. https://doi.org/10.1007/s40894-019-00115-z

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450. https://doi.org/10.1007/BF02294627

Watson, J. C. (2017). Establishing evidence for internal structure using exploratory factor analysis. *Measurement and Evaluation in Counseling and Development*, *50*(4), 232-238. https://doi.org/10.1080/07481756.2017.1336931

Wong, Z. Y., & Liem, G. A. D. (2022). Student engagement: Current state of the construct, conceptual refinement, and future research directions. *Educational Psychology Review*, *34*(1), 107-138. https://doi.org/10.1007/s10648-021-09628-3

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., & Goodman, N. (2020). Variational item response theory: Fast, accurate, and expressive. *arXiv 2002.00276*. https://doi.org/10.48550/arXiv.2002.00276

Yang, D., Wang, H., Metwally, A. H. S., & Huang, R. (2023). Student engagement during emergency remote teaching: A scoping review. *Smart Learning Environments*, *10*(1), Article 24. https://doi.org/10.1186/s40561-023-00240-2

Yin, Y., Shi, D., & Fairchild, A. J. (2023). The effect of model size on the root mean square error of approximation (RMSEA): The nonnormal case. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(3), 378-392. https://doi.org/10.1080/10705511.2022.2127729

Zhang, S., Chen, Y., & Liu, Y. (2018). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology*, *73*, 44-71. https://doi.org/10.1111/bmsp.12153

Zhang, Z., & McNamara, O. (2018). Key indicators of student engagement. In Z. Zhang & O. McNamara (Eds.), *Undergraduate student engagement: Theory and practice in China and the UK* (pp. 57-81). Springer. https://doi.org/10.1007/978-981-13-1721-7_4

APPENDIX
Additional Tables

Table 1A
Model fit for exploratory factor IRT models

Model	AIC	BIC	M2	*df*	RMSEA	SRMR	TLI	CFI
One factor	27386.14	27812.47	385.930	95	.068	.082	.805	.838
Two factors	26773.81	27280.92	237.424	77	.056	.049	.868	.911
Three factors	26647.14	27230.54	122.025	60	.040	.039	.934	.965

Note. AIC = Akaike information criteria; BIC = Bayesian information criteria; M2 = M2 statistic; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; TLI = Tucker-Lewis index; CFI = comparative fit index.

Table 2A
Likelihood ratio test model comparison

Model	*N* Par	logLik	χ^2	*df*	*p*
One factor	95	–13598.07	–	–	–
Two factors	113	–13273.91	648.329	18	0
Three factors	130	–13193.57	160.672	17	0

Note. *N* Par = number of parameters estimated; logLik = loglikelihood.

Table 3A
Standardized factor loading on one, two, and three dimensions

	F1	F2.1	F2.2	F3.1	F3.2	F3.3
i01	.74		.74		.71	
i02	.85		.86		.82	
i03	.82		.84		.80	
i04	.46		.41		.38	
i05	.88		.94		.94	
i06	.90		.94		.93	
i07	.87		.86		.84	
i08	.39	.37				.41
i09	.67	.31	.43		.36	.39
i10	.63	.31	.38		.31	.41
i11	.74	.47	.36			.62
i13	.54	.70				.52
i14	.66	.66				.76
i15	.54	.70		.41		.34
i16	.56	.66				.54
i17	.41	.64		.49		
i18	.59	.74		.85		
i19	.60	.80		.70		
i20	.64	.57		.63		

Note. F1 = loading factor for the unidimensional model; F2.1 = loading factor for the first dimension of the 2-dimensional model; F2.2 = loading factor for the second dimension of the 2-dimensional model; F3.1 = loading factor for the first dimension of the 3-dimensional model; F3.2 = loading factor for the second dimension of the 3-dimensional model; F3.3 = loading factor for the third dimension of 3-dimensional model.

Table 4A
Multidimensional discrimination and difficulty parameters

	MDISC	MDIFF1	MDIFF2	MDIFF3	MDIFF4
i01	2.00	–2.64	–1.57	–0.59	0.74
i02	2.98	–2.56	–1.57	–0.69	0.62
i03	2.60	–2.79	–1.75	–0.63	0.86
i04	0.90	–4.42	–3.04	–1.50	0.56
i05	3.97	–2.19	–1.44	–0.29	0.84
i06	4.35	–2.07	–1.30	–0.25	0.82
i07	3.19	–2.19	–1.45	–0.38	0.84
i08	0.84	–5.44	–4.57	–2.64	–0.49
i09	1.60	–2.89	–1.84	–0.13	1.37
i10	1.42	–2.75	–1.96	–0.57	0.84
i11	2.33	–2.72	–1.67	–0.07	1.39
i13	1.56	–3.26	–2.06	–0.29	1.32
i14	2.37	–2.95	–1.71	–0.33	1.21
i15	1.51	–2.75	–0.99	0.81	2.67
i16	1.66	–2.94	–1.37	0.62	2.34
i17	1.19	–3.82	–2.56	–0.89	1.35
i18	2.49	–2.40	–1.44	–0.15	1.38
i19	2.09	–3.67	–2.35	–0.88	0.99
i20	2.26	–2.52	–1.76	–0.44	1.12

Note. MDISC = multidimensional discrimination; MDIFF1 = multidimensional difficulty for choosing Category 2 (*disagree*); MDIFF2 = multidimensional difficulty for choosing Category 3 (*neither disagree nor agree*); MDIFF3 = multidimensional difficulty for choosing Category 4 (*agree*); MDIFF4 = multidimensional difficulty for choosing Category 5 (*strongly agree*).