Art. #2475, 12 pages, https://doi.org/10.15700/saje.v45ns2a2475

Psychometric properties of the Arabic MPVS-24 among Saudi Arabian high school students using item response theory

Hassan Alhaythami
Department of Psychology, College of Education, Umm Al-Qura University, Makkah, Saudi Arabia
hmali@uqu.edu.sa

Abstract
This study examined the psychometric properties of the Arabic version of the multidimensional peer victimisation scale (MPVS-24), using the graded response model of item response theory among high school students in Saudi Arabia. A quantitative method with a descriptive design was used to analyse the variables in this study. The sample size in this research was 641 students, selected randomly. For the data collected in the study the researcher used IRT for analysing the latent traits with precise measurement and identified item-level performance to further examine item discrimination and difficulty parameters. The results show acceptable values for item discrimination and trait thresholds, indicating good psychometric properties for the scale to measure bullying victimisation among high school students in Saudi Arabia. The Arabic MPVS-24 demonstrates strong psychometric properties, making it a reliable tool for assessing bullying victimisation among Saudi high school students. Its validated structure supports its use in future research and school-based interventions, with the potential for broader application in Arabic-speaking contexts.

Keywords: bullying; graded response model; high school; item response theory; human behaviour; psychometrics properties; Saudi students; victimisation

Introduction
Human behaviour refers to the actions and behaviour that individuals engage in to adapt to their environment and meet their needs (Rogers, 2020). External personal factors influence this behaviour. Some individuals' behaviour reflects their attempts to meet expectations and goals related to their age and level of education, and their behaviour may be driven by various needs that may not always align with social norms (Kalat, 2022). One example of this type of behaviour is bullying, which is socially rejected and can have negative effects on individuals and communities.

Bullying is a complex phenomenon that researchers define in various ways. Olweus (1996) defines bullying as aggressive behaviour that intentionally causes discomfort or harm to another person, while Batsche and Knoff (1994) describe it as aggressive behaviour in which there is a repeated imbalance of power or force. Nansel, Overpeck, Pilla, Ruan, Simons-Morton and Scheidt (2001) define bullying as a form of aggression in which an individual or group repeatedly attempts to harm weaker individuals. These definitions emphasise the deliberate, repeated, and harmful nature of bullying. This includes power differences between perpetrators and victims. Although the specifics may vary, all highlight the aggressive and dangerous nature of bullying, which includes the impact on victims.

Literature Review
One of the most prevalent forms is bullying in school. It involves unbalanced social interaction where verbal, physical, or electronic attacks are used to gain social control or domination (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2023). School bullying has negative consequences for all parties involved, including the perpetrator, victim, and school community. The school's security, psychological well-being, and social structure may be compromised by school bullying, and physical aggression towards bullies may harm students at any educational level. Numerous studies have shown that bullying has adverse effects on students' academic achievement and mental health. Victims of bullying are at an increased risk of mental health problems, academic difficulties, and social isolation, with some studies indicating a link to depression, anxiety, and suicidality (Al-Rahmi, Yahaya, Alamri, Aljarboa, Kamin & Saud, 2019; Kann, McManus, Harris, Shanklin, Flint, Queen, Lowry, Chyen, Whittle, Thornton, Lim, Bradford, Yamakawa, Leon, Brener & Ethier, 2018; Samara, Da Silva Nascimento, El-Asam, Hammuda & Khattab, 2021). Perpetrators of bullying are also at risk of negative outcomes, such as lower academic achievement and engagement, and a higher likelihood of engaging in other forms of aggressive behaviour (Ngidi & Moletsane, 2018; Olweus, 1996; Sourander, Gyllenberg, Brunstein Klomek, Sillanmäki, Ilola & Kumpulainen, 2016).

School bullying is considered one of the most significant social problems in Saudi Arabia, which has garnered the attention of researchers in the field of social life. According to the global study on Trends in the International Mathematics and Science Study ([TIMSS], Mullis, Martin, Foy & Arora, 2012), 40% of students in Saudi schools have experienced school bullying (Richardson & Hiu, 2018). As a result, school bullying has become a matter of concern for authorities and staff at education institutions who work to identify the extent of the problem and apply measures to reduce it (Smokowski & Kopasz, 2005). These include preventive programmes and laws implemented in schools to punish the bullies (The National Family Safety Program,

2016). The Saudi Arabian Ministry of Education has recognised the enormity of the problem in schools and considers it to be a national issue. However, despite this recognition, studies on the fast-spreading problem of school bullying in the Saudi school community have still not received the necessary attention (Al-Sharif, 2018; Al-Zahrani, 2020; Khouj, 2012).

Many researchers are interested in developing measures with good psychometric properties and accurately capture the problem of bullying in schools, including solutions to eliminate or reduce impacts on the education system. As a result, several bullying scales have been developed with unique sets of items and scoring criteria. One of the most widely used questionnaires is the bully/victim questionnaire, developed by Olweus in 1996. The bullying and victimisation questionnaire (BVQ) includes 36 items that assess both the frequency and severity of bullying behaviour, as well as victims' reactions to bullying. It has demonstrated good reliability and validity and is widely used in both research and clinical settings (Olweus, 1996).

Another level of bullying measurement is the adolescent peer relationships instrument (APRI). The APRI is a self-report questionnaire developed by Parada (2000) with the purpose of measuring various dimensions of adolescent peer relationships. The APRI consists of 36 items and measures various dimensions of peer relationship, including acceptance/rejection from peers, popularity, social support, loneliness, and victimisation. The APRI shows good psychometric properties, including high reliability and validity. Several studies have attempted to examine the role of peer relationships in adolescents' adjustment and well-being (Parada, 2000).

The MPVS-24 is one of the most popular measures of peer victimisation. The original version developed by Mynard and Joseph (2000) included 16 items assigned to assess four types of victimisation: physical, verbal, social manipulation and attacks on property. This scale was designed to measure the frequency and severity of these experiences.

However, it was difficult to find scales with good psychometric properties in Arabic that may be used to measure bullying or its victims. Due to this challenge, the researcher decided to address this gap in the literature by evaluating the Arabic version of the MPVS-24 scale to measure bullying victimisation. The goal of this research was to contribute to the development of effective tools to measure bullying victimisation in the Arabic-speaking population, and ultimately, to help reduce the prevalence of this dangerous phenomenon.

Item response theory (IRT)

Psychologists have been actively involved in the measurement movement since the development of psychometric tools aimed at improving their objectivity and reliability. According to classical testing theory (CTT), an individual's observed score consists of the true score and the error. This theory suggests that an individual's abilities change over time. It is influenced by the difficulty or ease of the test items. In the same way, test items change in properties according to individual characteristics. This reveals a reciprocal relationship between the test and the performance of the person taking the test (Hogan, 2014; Lord & Novick, 1968). Reliability, a key focus of CTT, measures the consistency of test scores. Researchers such as Cronbach (1951) have introduced methods such as Cronbach's alpha to assess the internal organisation of test items to ensure accuracy in measuring the desired structure.

This statistical foundation underpins the design of reliable psychometric assessment across disciplines (Cronbach, 1951; Dunn, Baguley & Brunsden, 2014). As the fields of education and psychology have advanced, it has become evident that using tests alone to assess an individual's abilities and psychological traits is inadequate. As a result, traditional measurement theory has come under criticism for its failure to achieve measurement objectivity, as it relies on the properties of the test, such as its validity, reliability, and discrimination. Therefore, results may vary depending on the available tests. This is coupled with the nature of the sample population being measured (Hambleton & Swaminathan, 2013).

Not only does this encourage researchers and measurement enthusiasts to seek new approaches, but their efforts have also resulted in modern trends in the field of measurement. Along with the emergence of a new type of theory called latent trait theory or general object response theory, in psychometrics in the 1950s and 1960s, Frederick Lord and others developed IRT to create an instrument that scores respondents on the same test (Hambleton & Jodoin, 2003; Lord, 1952; Lord & Novick, 2008). Many of the problems originally associated with measurement have been solved. IRT is a statistical model used to infer the abilities or latent characteristics of individuals. It is based on responses to a dataset (Hambleton & Swaminathan, 2013) and is widely used in educational and psychological assessment as well as market research.

This means that a key advantage of IRT over classical test theory is the separate analysis of each test item. Instead, it is an important part of the whole (De Ayala, 2013), so the finer features of the test can be assessed. This includes the level of difficulty and the power of the programme in separating people with different abilities. This information can be further used to improve the quality, fairness, and reliability of trials. Another strength of IRT lies in its application to partial

credit items, that is, items that give partial credit instead of binary responses (correct/incorrect) which is different from classic testing. Consequently, measurement results could be different based on the test used, besides the characteristics of the sample population being measured (Hambleton & Swaminathan, 2013).

Graded response model

The graded response model (GRM) is a type of advanced IRT model, commonly used in psychometrics for analysing data in which the items are categorical and ordinal, such as most Likert-scale items. Assume that for each item a set of threshold parameters exists – points on the underlying latent trait continuum where individuals are equally likely to respond in adjacent categories (Samejima, 1969). The GRM enables one to estimate item and person parameters, whereby it is assumed that the probability of a response to a certain item depends on, among others, the location of an individual on the latent trait measured, together with the difficulty and discrimination parameters of the item itself. It makes this especially useful for data from surveys or questionnaires where items are scored in graded scales. The GRM has been in vogue for the development and validation of psychometric scales, as it provides more appropriate estimates of the latent trait under consideration than the classical test theory methods. It has been most frequently used in research into diverse topics in the areas of personality, attitudes, and mental health.

The Purpose of the Study

Despite the advancements in science and the increase in knowledge in recent years, bullying remains a widespread problem in modern society. It manifests in various forms in schools, workplaces, and communities, and its nature can vary depending on location, time, and individual differences, as well as how individuals experience or perpetrate bullying (UNESCO, 2023). Since Saudi schools suffer greatly from various types of bullying (Al-Sharif, 2018; Al-Zahrani, 2020; Khouj, 2012; Mullis et al., 2012), the researcher was interested in determining a valid Arabic scale that may be used for these types of students, which reflects some degree of bullying. This study was conducted in order to assist supervisors and educators in reducing bullying and its impact; the study sought to answer the main research question: What are the psychometric properties of the Arabic version of the multidimensional peer victimisation scale (MPVS-24) among high school students in Saudi Arabia using item response theory?

Method

The Study Design

This study used a quantitative, cross-sectional design to evaluate the psychometric properties of the multidimensional victimisation scale (MPVS) in Saudi high school students. The focus of the design was on evaluating the reliability, validity, and dimensionality of the scale using IRT. Data collection involved a large, representative sample of high school students and rigorous statistical analysis was performed to assess item heterogeneity and measurement validity. The design allowed for an in-depth exploration of peer victimisation experiences, even though it was once-off with no long-term follow-up.

Population and Sample

In this study, a survey was conducted among students in 98 high schools in the Al-Qunfudhah province of Saudi Arabia. The Department of Education in the province helped distribute the survey, which was administered in Arabic after being translated from English. Of the 13,085 secondary school students in the area, 641 students were selected randomly to complete the questionnaire. The respondents, drawn from three different grades, were 358 (55.9%) male and 283 (44.1%) female.

Instrument

The MPVS was developed by Mynard and Joseph (2000) with the MPVS-24 as an updated version of the scale (Joseph & Stockton, 2018). The original MPVS consisted of a 16-item scale. Four peer categories measured experiences of victimisation: physical, verbal, social manipulation and attacks on property. The MPVS-24 contains 24 items measuring the frequency and severity of six categories of victimisation experiences, whereas the original MPVS measured four categories. Eight items were added to the MPVS-24 to assess experiences of electronic victimisation and social rejection.

The MPVS has been used in many studies and has demonstrated good psychometric properties, including reliability and validity. Many researchers found it to be a useful tool for identifying experiences of peer victimisation in children and adolescents (Vessey, Strout, DiFazio & Walker, 2014; Vivolo-Kantor, Martell, Holland & Westby, 2014). The MPVS-24 is considered a more comprehensive and detailed measure of peer victimisation experiences. This makes it a very useful tool for researchers and practitioners who want to understand and deal with bullying and victimisation among children and adolescents. Since this is considered an improved scale, not much information about its psychometric properties is available. The scale was translated into Arabic and all items were measured using a 3-point Likert response scale (0 = Not at all, 1 = Once, 2 = More than once). In addition, the wording for some items was changed, and some items were replaced with new items to make them more suitable for the

cultural context of the study population. The response scale of the Arabic version of MPVS-24 was also changed to a 5-point Likert scale (1 = Not at all, 2 = Rarely, 3 = Sometimes, 4 = Often, and 5 = Very often (see Appendix A).

Procedure

Prior to conducting the study, approval for using and translating the scale was obtained from the developer of the scale. Thereafter, the researcher first translated the original English version of the MPVS-24 into Arabic. Then, two native Arabic speakers fluent in English translated the English measure into Arabic. The Arabic version of the MPVS-24 was translated back into English by two English-speaking language specialists who were unaware of the original scale. The differences between the translated versions were examined to ensure agreement between the two versions. Before distributing the scale to high school students, approval was obtained from the Department of Education in the Al-Qunfudah province.

Statistical Analysis

After completing the data collection process, the data were converted to an Excel file for cleaning and preparation for statistical processing. The Statistical Package for the Social Sciences (SPSS) program was used to analyse the characteristics of the study sample and verify the assumption of unidimensionality through exploratory factor analysis. IRTPRO software was then used to process the main data and estimate the difficulty, discrimination, and information function for the items on the Arabic scale for adolescent peer relations according to the GRM, a modern theory model suitable for data extracted from a Likert scale response.

Results

Verifying the Assumptions of Graded Response Model

The assumption of unidimensionality

The unidimensionality assumption means that a test or measure assesses only one underlying construct or dimension. In other words, it assumes that all items in the test measure the same construct and that the scores on the test are determined by a single latent trait. After verifying that the data were appropriate for conducting an exploratory factor analysis, the analysis was carried out. Based on the results of the exploratory factor analysis (EFA), the data show a total variance of 36.17%, indicating that the factor structure was moderately complex. The scree plot also suggests a break after the fourth factor, which suggests that a four-factor solution may be appropriate. The first factor explained 26.93% of the variance, while the second, third, and fourth factors explained 3.14%, 3.08%, and 3.02% of the variance, respectively (see Table 1). The remaining factors explained less than 3% of the variance each. According to Hattie (1985), if the percentage of variance explained by the first factor is greater than 20%, it can be inferred that the test or measure is unidimensional. Based on this, the results suggest that the assumption of unidimensionality has been met for this dataset.

Table 1 Results of factor analysis

Component	Initial eigenvalues	% of variance	Cumulative %
1	9.696	26.93	26.93
2	1.129	3.14	30.07
3	1.109	3.08	33.15
4	1.087	3.02	36.17
5	1.036	2.88	39.05

The assumption of local independence

The local independence in IRT means that an individual's response to a particular item is not influenced by their responses to other items, given that their ability level on the construct being measured is held constant (Hambleton, Swaminathan & Rogers, 1991). This implies that the items on a test are independent of one another and do not interact with each other in determining the test scores. Based on the results, the range of values obtained from the χ^2 LD (linkage disequilibrium) statistic for all pairs of items in the test was between -1.5 to 7.1, which falls below the recommended threshold value of 10. Therefore, it can be concluded that the test items were sufficiently locally independent to apply the GRM for analysing the items.

Estimating Item Parameters for the Graded Response Model

Overall, the fit indices for the GRM indicated that it provided an adequate fit to the overall dataset (M_2 = 334.63, df = 180, p = 0.0001, RMSEA = 0.04). However, upon examining each item's fit to the model, the results indicate that several items had poor fit, as shown by significant values for the S-χ^2 fit index. Specifically, Item 1 ("Punched me", χ^2 = 166.79, df = 135, p = 0.03), Item 9 ("Hurt me physically in some way", χ^2 = 167.49, df = 138, p = 0.04), Item 12 ("Stole something from me", χ^2 = 170.16, df = 136, p = 0.02), Item 13 ("Beat me up", χ^2 = 176.66, df = 140, p = 0.01), Item 17 ("Sent me a nasty text", χ^2 = 174.19, df = 144, p = 0.04), and Item 23 ("Spread rumours about me among friends", χ^2 = 191.91, df = 141, p = 0.002) had significant S-χ^2 values, indicating poor fit to the IRT model with reference. These results suggest that these items may need to be revised or removed from the scale in order to improve the overall fit of the model. As is recommended by Toland (2014), these items (1, 9, 12, 13, 17, and 23) were removed from the dataset before proceeding with further analysis to ensure the accuracy of the analysis. A summary of the item-level fit results is presented in Table 2.

Table 2 Item-level fit statistics based on S-X^2 test

Item	X^2	*p*	Item	X^2	*p*
1	166.79	0.03	13	176.66	0.01
2	159.16	0.07	14	144.95	0.34
3	164.36	0.05	15	143.68	0.51
4	157.03	0.19	16	158.98	0.12
5	163.32	0.10	17	174.19	0.04
6	146.77	0.23	18	139.90	0.48
7	154.58	0.18	19	167.44	0.07
8	154.86	0.15	20	155.17	0.19
9	167.49	0.04	21	140.01	0.41
10	159.40	0.09	22	153.96	0.10
11	167.09	0.05	23	191.91	0.002
12	170.16	0.02	24	146.35	0.40

The item parameters that were estimated using the GRM within the framework of IRT for the items of the Arabic version of MPVS are presented in Table 3. The GRM estimates a slope parameter for each response option, which indicates the steepness of the curve for that option, as well as threshold parameters for each response category, which indicate the location of the curve along the latent trait axis. The discrimination parameters (a) ranged from 0.44 to 0.91. These values indicate that some items were more effective than others in discriminating between examinees at different levels of the latent trait. Item 22 ("Would not let me join in their game") had the highest discrimination parameter (a = 0.91, *SE* = 0.12), while Item 4 ("Took something of mine without permission") had the lowest discrimination parameter (a = 0.44, *SE* = 0.09).

The threshold parameters (b) represent the point along the latent trait axis at which an examinee had a 50% chance of selecting a particular response category. The threshold parameters ranged from -3.78 to -2.35 for the first category, -1.78 to -.58 for the second category, -0.10 to 1.68 for the third category, and 1.11 to 3.77 for the fourth category. These values suggest that the items varied in their level of difficulty, with some items being easier to endorse than others. For example, Item 4 ("Took something of mine without permission") had the lowest category threshold parameter (b_1 = -3.78, *SE* = 0.77), indicating that it was the easiest item in the set, while Item 5 ("Made fun of me because of my appearance") had the highest category threshold parameter (b_4 = 3.77, *SE* = .73), indicating that it was the most difficult item in the set. Overall, the results suggest that the items in this set varied in their effectiveness in discriminating between examinees at different levels of the latent trait, as well as in their level of difficulty. These findings have implications for the measurement of the construct of interest and the interpretation of scores on this measure.

Table 3 Graded response model parameter estimates for the Arabic version of MPVS-24

Item	a (*SE*)		b_1 (*SE*)		b_2 (*SE*)		b_3 (*SE*)		b_4 (*SE*)	
2	0.69	(0.10)	-2.77	(0.38)	-1.08	(0.18)	0.49	(0.14)	2.34	(0.33)
3	0.63	(0.09)	-2.56	(0.38)	-0.85	(0.17)	0.91	(0.18)	2.59	(0.39)
4	0.44	(0.09)	-3.78	(0.77)	-1.67	(0.37)	0.97	(0.27)	3.27	(0.67)
5	0.47	(0.09)	-3.04	(0.59)	-0.58	(0.20)	1.68	(0.36)	3.77	(0.73)
6	0.81	(0.10)	-2.35	(0.29)	-0.99	(0.15)	0.39	(0.11)	1.99	(0.26)
7	0.61	(0.10)	-3.40	(0.53)	-1.52	(0.26)	0.29	(0.14)	2.25	(0.36)
8	0.54	(0.09)	-3.54	(0.60)	-1.42	(0.27)	0.45	(0.17)	2.76	(0.47)
10	0.57	(0.09)	-2.93	(0.47)	-1.22	(0.23)	0.71	(0.18)	2.77	(0.45)
11	0.67	(0.10)	-3.10	(0.45)	-1.62	(0.25)	-0.10	(0.12)	1.55	(0.25)
14	0.70	(0.10)	-2.60	(0.36)	-1.25	(0.20)	0.70	(0.15)	2.46	(0.35)
15	0.50	(0.09)	-3.18	(0.58)	-0.82	(0.21)	1.12	(0.25)	3.55	(0.64)
16	0.55	(0.09)	-3.34	(0.55)	-1.55	(0.28)	0.75	(0.19)	2.98	(0.50)
18	0.73	(0.10)	-2.61	(0.35)	-1.07	(0.17)	0.68	(0.14)	2.28	(0.31)
19	0.52	(0.09)	-3.18	(0.56)	-1.30	(0.27)	0.71	(0.20)	2.92	(0.52)
20	0.59	(0.09)	-2.99	(0.48)	-1.26	(0.23)	0.82	(0.19)	2.71	(0.44)
21	0.69	(0.10)	-2.92	(0.41)	-1.42	(0.22)	0.21	(0.12)	1.96	(0.29)
22	0.91	(0.12)	-2.40	(0.29)	-1.20	(0.16)	0.02	(0.09)	1.11	(0.16)
24	0.56	(0.09)	-3.49	(0.58)	-1.78	(0.32)	0.08	(0.15)	2.19	(0.38)

Note. a refers to the slope parameter; b_1, … b_4 refer to the four threshold parameters; *SE* refers to the standard error.

Item Information Function and Test

Information about the item information function (IIF) and test information function (TIF) of the GRM is shown in Table 4. The IIF represents the amount of information provided by each item at different levels of the latent trait (θ), while the TIF represents the amount of information provided by the entire test at different levels of the latent trait. The table shows the values of the IIF and TIF for 14 items at different levels of θ, ranging from -2.8 to 2.8, in increments of 0.4. The IIF values range from 0.11 to 0.29, with higher values indicating more information provided by the item. For example, Item 8 provides the most information at θ = 0.0, with an IIF value of 0.29, while item 14 provides the least information at θ = -2.8, with an IIF value of 0.11. In general, the IIF values tend to be highest at the mid-range of the latent trait (around θ = 0.0) and lower at the extreme ends (θ = -2.8 and 2.8).

The TIF values in the table range from 1.30 to 5.65, with higher values indicating more information provided by the entire test. The TIF values are highest at the mid-range of the latent trait (around $\theta = 0.0$) and lower at the extreme ends ($\theta = -2.8$ and 2.8). Overall, the information provided by each item and the entire test varies depending on the level of the latent trait (see Figure 1). The IIF and TIF values can be used to evaluate the effectiveness of each item and the entire test in measuring the latent trait of interest.

Table 4 Graded response the model item information function and the test information function

Item	θ	-2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	-0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
2		0.14	0.15	0.16	0.17	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.17	0.17
3		0.16	0.19	0.21	0.22	0.23	0.24	0.24	0.25	0.25	0.25	0.25	0.24	0.24	0.23	0.22
4		0.17	0.20	0.22	0.24	0.25	0.25	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.24	0.22
5		0.14	0.16	0.18	0.19	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.20	0.19
6		0.15	0.18	0.20	0.21	0.22	0.23	0.23	0.24	0.24	0.24	0.24	0.24	0.23	0.22	0.21
7		0.15	0.17	0.19	0.20	0.21	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.21	0.20
8		0.17	0.21	0.24	0.26	0.27	0.28	0.29	0.29	0.29	0.29	0.29	0.29	0.28	0.27	0.24
10		0.13	0.14	0.15	0.16	0.16	0.16	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.16	0.16
11		0.13	0.15	0.16	0.16	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.16
14		0.11	0.12	0.13	0.13	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
15		0.17	0.20	0.22	0.23	0.24	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.24	0.24	0.22
16		0.12	0.14	0.15	0.15	0.16	0.16	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.16	0.16
18		0.14	0.15	0.16	0.17	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.19	0.18	0.18	0.17
19		0.15	0.17	0.19	0.21	0.22	0.22	0.22	0.23	0.23	0.23	0.23	0.23	0.22	0.21	0.20
20		0.14	0.15	0.16	0.17	0.17	0.17	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.17	0.17
21		0.14	0.16	0.18	0.19	0.20	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.20	0.19
22		0.16	0.19	0.22	0.24	0.25	0.26	0.26	0.27	0.27	0.27	0.27	0.26	0.26	0.24	0.22
24		0.13	0.14	0.15	0.16	0.17	0.17	0.17	0.18	0.18	0.18	0.18	0.18	0.17	0.17	0.17
Test		3.59	3.95	4.24	4.46	4.61	4.70	4.75	4.78	4.80	4.81	4.80	4.77	4.71	4.59	4.40
SE		0.53	0.50	0.49	0.47	0.47	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.47	0.48

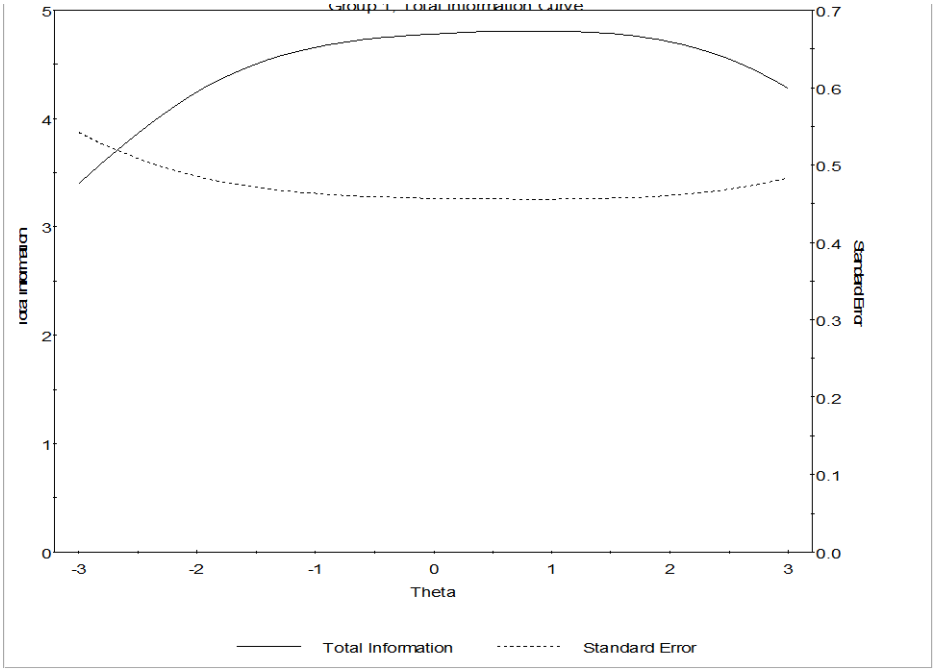

Figure 1 The test information function of the Arabic version of MPVS-24 (graded response model)
Note. Theta = level of empathic behaviour on the latent trait continuum.

Discussion

Bullying is a huge problem affecting scores of people in the world, including in Saudi Arabia. Research in bullying, however, has increased lately, but most of this involves assessment of bullying in other countries and does not assess the nature of bullying in an Arabic-speaking country like Saudi Arabia. This, therefore, poses a challenge for researchers and practitioners who seek to measure the prevalence and impact of bullying accurately in the Saudi context. Having a level of harassment in Arabic is especially important in the Saudi context. For example, because of cultural and linguistic differences, some behaviour that are considered bullying in Western culture may not be regarded as such in Saudi Arabia. Similarly, cultural differences may affect the perception of bullying and its response. Therefore, it becomes difficult to apply the measures developed in a different cultural context.

GRM is a type of IRT used in the psychometric analysis of multiple measures, such as surveys or questionnaires. In this study the researcher used GRM to explore the psychometric properties of the Arabic version of the MPVS-24.

This study reveals that, in general, GRM provided a good fit to the data. This is indicated by various fit indices such as M_2 and RMSEA. However, upon reviewing each of the items separately, it was found that quite a number of the models had poor fit as shown by significant values of the S-χ^2 fit index, especially for Items 1, 9, 12, 13, 17, and 23 as having significant S-χ^2 values. By reference to IRT, it is demonstrated that the model is not fit. These items were, therefore, removed from the data set before further analysis.

In addition, the items differ in their discrimination parameter (a) and threshold parameter (b). The discrimination threshold indicates how effectively an item is in discriminating between test subjects at a given latent symptom level. A different higher value indicates greater discrimination. The rubric specifies the level of difficulty for each item. A lower value indicates that the item is easier and higher values indicate more difficulty. Overall, scale items covered a wide range of latent symptoms -3.78 (Item 4, b_1) to 3.77 (Item 5, b_4). The parameter reflects the ability of the response category to discriminate between individuals on a scale. The different measured characteristics ranged from .44 to .91, indicating moderate levels of discrimination (Baker, 2001).

Furthermore, the study examined the IIF and TIF of GRM. IIF represents the amount of information provided by each item in different levels of latent symptoms, while TIF represents the amount of information provided by each item by examining all the underlying symptoms at various levels.

Overall, the findings of this study have important implications for the measurement of peer victimisation and the interpretation of scores on the MPVS. This study provides valuable information for improving the validity of the MPVS. It measures peer victimisation by identifying offending items and examining psychometric properties.

These findings strengthen the MPVS as a reliable and valid tool for assessing peer victimisation. It is suitable for a variety of research and practical contexts. Its psychometric robustness has been demonstrated through studies to improve items and its validity in various populations (Balogun & Olapegba, 2007; Betts, Houston & Steer, 2015; Morrow, Hubbard & Swift, 2014). This makes the MPVS an important tool for measuring peer victimisation that is precise and meaningful. It allows for application in a variety of educational and psychological settings.

Conclusion

This study examined the psychometric properties of the multidimensional peer victimisation scale (MPVS) of Saudi high school students using IRT and provide insight into the psychometric properties of the MPVS indicator in Saudi high school students, measuring peer victimisation experiences. In the case of Saudi Arabia, there is great value when it comes to rigorous analysis and investigation.

The findings from this study demonstrate that the MPVS has adequate psychometric properties. Including good item discrimination and measurement validity, IRT improves understanding of performance at different levels. Therefore, there is appropriate assessment and interpretation of experiences related to peer victimisation among high school students in Saudi Arabia. The study also highlights the multidimensionality of peer victimisation by detailing various dimensions and influences that have a specific impact on students' well-being and social adjustment. These may help design specific intervention and prevention strategies to address unique patterns of victimisation among Saudi high school students.

Overall, this study has significantly enhanced our understanding of the psychometric properties and practical utility of the MPVS in capturing the multifaceted nature of peer victimisation among Saudi high school students. Furthermore, the study offers valuable insight for educators, researchers, and policymakers in developing effective strategies to prevent and mitigate the negative impacts of peer victimisation within the Saudi educational context.

The study acknowledges several limitations that should be addressed in future research. The sample was limited to Saudi high school students. It limits the normalisation of conclusions for other age groups or cultural contexts. Although the use of the IRT provides valuable insight, this study did not compare the results with other outlines, such as CTT, and cross-sectional design, preventing conclusions about the stability of MPVS over time. In future research MPVS in diverse populations and cultural contexts should be examined to increase its applicability. Taking a long-term perspective may provide insight into the stability of peer victimisation experiences and the effectiveness of their levels over time. Comparative studies involving different psychometric frameworks would further confirm its robustness. Moreover, exploring targeted interventions for identified dimensions of peer victimisation may be of practical benefit for educators and policymakers to address these issues.

Note

References

Al-Rahmi WM, Yahaya N, Alamri MM, Aljarboa NA, Kamin YB & Saud MSB 2019. How cyber stalking and cyber bullying affect students' open learning. *IEEE Access*, 7:20199–20210. https://doi.org/10.1109/access.2019.2891853

Al-Sharif I 2018. The role of the school administration in addressing the phenomenon of school bullying in the intermediate stage from one point of view of male and female students in Jeddah. *Journal of the College of Education*, 34(3).

Al-Zahrani SA 2020. The efforts of the Kingdom of Saudi Arabia to confront the problem of school bullying in the light of the forces and factors affecting it - A comparative study. *Journal of Educational and Psychological Sciences*, 4(36):163–183. https://doi.org/10.26389/ajsrp.s130420

Baker FB 2001. *The basics of item response theory* (2nd ed). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Available at https://files.eric.ed.gov/fulltext/ED458219.pdf. Accessed 31 December 2025.

Balogun SK & Olapegba PO 2007. Cultural validation of the Multidimensional Peer Victimization Scale in Nigerian children. *Journal of Cross-Cultural Psychology*, 38(5):573–580. https://doi.org/10.1177/0022022107305239

Batsche GM & Knoff HM 1994. Bullies and their victims: Understanding a pervasive problem in the schools. *School Psychology Review*, 23(2):165–174. https://doi.org/10.1080/02796015.1994.12085704

Betts LR, Houston JE & Steer OL 2015. Development of the Multidimensional Peer Victimization Scale–Revised (MPVS-R) and the Multidimensional Peer Bullying Scale (MPVS-RB). *The Journal of Genetic Psychology*, 176(2):93–109. https://doi.org/10.1080/00221325.2015.1007915

Cronbach LJ 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334. https://doi.org/10.1007/BF02310555

De Ayala RJ 2013. *The theory and practice of item response theory*. New York, NY: Guilford Publications.

Dunn TJ, Baguley T & Brunsden V 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3):399–412. https://doi.org/10.1111/bjop.12046

Hambleton RK & Jodoin M 2003. Item response theory: Models and features. In R Fernández-Ballesteros (ed). *Encyclopedia of psychological assessment*

(Vol. 1). London, England: Sage. Available at https://arthurjensen.net/wp-content/uploads/2014/06/2003-fernandezballesteros-encyclopediaofpsychologicalassessment.pdf. Accessed 31 December 2025.

Hambleton RK & Swaminathan H 2013. *Item response theory: Principles and applications*. New York, NY: Springer Science+Business Media.

Hambleton RK, Swaminathan H & Rogers HJ 1991. *Fundamentals of item response theory* (Vol. 2). Newbury Park, CA: Sage.

Hattie J 1985. Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2):139–164. https://doi.org/10.1177/014662168500900204

Hogan TP 2014. *Psychological testing: A practical introduction* (3rd ed). Hoboken, NJ: John Wiley & Sons.

Joseph S & Stockton H 2018. The multidimensional peer victimization scale: A systematic review. *Aggression and Violent Behavior*, 42:96–114. https://doi.org/10.1016/j.avb.2018.07.009

Kalat JW 2022. *Introduction to psychology* (12th ed). Boston, MA: Cengage Learning.

Kann L, McManus T, Harris WA, Shanklin SL, Flint KH, Queen B, Lowry, R, Chyen D, Whittle L, Thornton J, Lim C, Bradford D, Yamakawa Y, Leon M, Brener N & Ethier KA 2018. Youth risk behaviour surveillance — United States, 2017. *MMWR Surveillance Summaries*, 67(8):1–114. https://doi.org/10.15585/mmwr.ss6708a1

Khouj HA 2012. School bullying and its relationship to social skills in primary school students in Jeddah, Saudi Arabia. *Journal of Educational & Psychological Sciences*, 13(4):187–218. https://doi.org/10.12785/JEPS/130407

Lord F 1952. *A theory of test scores* (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation. Available at https://www.psychometricsociety.org/sites/main/files/file-attachments/mn07.pdf?1576607452. Accessed 31 December 2025.

Lord FM & Novick MR 1968. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing.

Lord FM & Novick MR 2008. *Statistical theories of mental test scores*. Charlotte, NC: Information Age Publishing.

Morrow MT, Hubbard JA & Swift LE 2014. Relations among multiple types of peer victimization, reactivity to peer victimization, and academic achievement in fifth-grade boys and girls. *Merrill-Palmer Quarterly*, 60(3):302–327. https://doi.org/10.13110/merrpalmquar1982.60.3.0302

Mullis IVS, Martin MO, Foy P & Arora A 2012. *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Available at https://timssandpirls.bc.edu/timss2011/downloads/t11_ir_mathematics_fullbook.pdf. Accessed 31 December 2025.

Mynard H & Joseph S 2000. Development of the multidimensional peer-victimization scale. *Aggressive Behaviour*, 26(2):169–178. https://doi.org/10.1002/(SICI)1098-2337(2000)26:2%3C169::AID-AB3%3E3.0.CO;2-A

Nansel TR, Overpeck M, Pilla RS, Ruan WJ, Simons-Morton B & Scheidt P 2001. Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment. *JAMA*, 285(16):2094–2100. https://doi.org/10.1001/jama.285.16.2094

Ngidi ND & Moletsane R 2018. Bullying in school toilets: Experiences of secondary school learners in a South African township [Special issue]. *South African Journal of Education*, 38(Suppl. 1):Art. #1588, 8 pages. https://doi.org/10.15700/saje.v38ns1a1588

Olweus D 1996. *The revised Olweus bully/victim questionnaire*. Bergen, Norway: Research Center for Health Promotion.

Parada R 2000. *Adolescent Peer Relations Instrument: A theoretical and empirical basis for the measurement of participant roles in bullying and victimisation of adolescence: An interim test manual and a research monograph: A test manual.* Sydney, Australia: Publication Unit, Self-concept Enhancement and Learning Facilitation (SELF) Research Centre, University of Western Sydney.

Richardson D & Hiu CF 2018. *Developing a global indicator on bullying of school-aged children*. Florence, Italy: UNICEF Office of Research – Innocenti.

Rogers A 2020. *Human behaviour in the social environment: Perspectives on development, the life course, and macro contexts* (2nd ed). New York, NY: Routledge. https://doi.org/10.4324/9781003025382

Samara M, Da Silva Nascimento B, El-Asam A, Hammuda S & Khattab N 2021. How can bullying victimisation lead to lower academic achievement? A systematic review and meta-analysis of the mediating role of cognitive-motivational factors [Special issue]. *International Journal of Environmental Research and Public Health*, 18(5):2209. https://doi.org/10.3390/ijerph18052209

Samejima F 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1):1–97. https://doi.org/10.1007/bf03372160

Smokowski PR & Kopasz KH 2005. Bullying in school: An overview of types, effects, family characteristics, and intervention strategies. *Children & Schools*, 27(2):101–110. https://doi.org/10.1093/cs/27.2.101

Sourander A, Gyllenberg D, Brunstein Klomek A, Sillanmäki L, Ilola AM & Kumpulainen K 2016. Association of bullying behavior at 8 years of age and use of specialized services for psychiatric disorders by 29 years of age. *JAMA Psychiatry*, 73(2):159–165. https://doi.org/10.1001/jamapsychiatry.2015.2419

The National Family Safety Program 2016. *"Training guide, awareness program for the prevention of domestic violence."* Awareness Department - Department of Community Services.

Toland MD 2014. Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, 34(1):120–151.

https://doi.org/10.1177/0272431613511332
United Nations Educational, Scientific and Cultural Organization 2023. *Global Education Monitoring Report Summary 2023: Technology in education: A tool on whose terms?* Paris, France: Author. Available at https://www.unesco.org/gem-report/sites/default/files/medias/fichiers/2023/07/Summary_v5.pdf. Accessed 31 December 2025.
Vessey J, Strout TD, DiFazio RL & Walker A 2014. Measuring the youth bullying experience: A systematic review of the psychometric properties of available instruments. *Journal of School Health*, 84(12):819–843. https://doi.org/10.1111/josh.12210
Vivolo-Kantor AM, Martell BN, Holland KM & Westby R 2014. A systematic review and content analysis of bullying and cyber-bullying measurement strategies. *Aggression and Violent Behavior*, 19(4):423–434. https://doi.org/10.1016/j.avb.2014.06.008

Appendix A: The Arabic Version of the MPVS-24 in English

How often during the last school year has another pupil done these things to you? Please answer by putting a tick in one of the five columns for each of the 24 questions.

1) Not at all (0–1 time per month), 2) Rarely (2–3 times per month) 3) Sometimes (4–6 times per month), 4) Often (7–10 times per month), and 5) Very often (more than 10 times per month).

Item	Not at all	Rarely	Sometimes	Often	Very often
Punched me					
Tried to get me into trouble with my friends					
Called me bad names*					
Took something of mine without permission					
Kicked me					
Tried to make my friends turn against me					
Made fun of me because of my appearance					
Tried to break something of mine					
Hurt me physically in some way					
When I tried to play with one person, another person would not let me					
Made fun of me for some reason					
Stole something from me					
Beat me up					
Made other people not talk to me					
Swore at me					
Deliberately damaged some property of mine					
Sent me a nasty text					
Ignored me					
Said something mean about me on a social networking site					
Refused to talk to me					
Said spiteful things about me in a classroom*					
Left me out of activities or games on purpose					
Spread rumours about me among friends**					
Had a secret and would not tell me					

Note. *Some changes on the wording of the MPVS-24, **New items in the Arabic Version of MPVS-24.