British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 3, 2025 (pp. 72-85)
www.abjournals.org

USE OF ITEM RESPONSE THEORY TO VALIDATE EMOTIONAL VACILLATION SCALE FOR UNDERGRADUATES IN NIGERIA

Godwin Matthew Sabboh

Educational Research Methodology, School of Education,
University of North Carolina at Greensboro, USA.

Email: gmsabboh@uncg.edu

Cite this article:

Godwin Matthew Sabboh (2025), Use of Item Response Theory to Validate Emotional Vacillation Scale for Undergraduates in Nigeria. British Journal of Education, Learning and Development Psychology 8(3), 72-85. DOI: 10.52589/BJELDP-6LKXN148

Manuscript History

Received: 11 Oct 2025

Accepted: 4 Nov 2025

Published: 18 Nov 2025

ABSTRACT: *A ten-item emotional vacillation scale (EVS) was coined from the literature of Sullivan and Strongman (2003) on mixed emotion and was adapted by the author to measure how emotion changes and/or fluctuate rapidly and its levels among undergraduates. Unlike other similar concepts, such as emotional regulation, there are no scales developed to measure emotional vacillation among undergraduates. A cross-sectional research design was adopted and a non-probability sampling technique known as convenience sampling was used. The sample size comprised five-hundred and sixty-one (N =561) undergraduates in Nigerian Universities. Also, psychometric evidence was obtained to evaluate whether it is appropriate to use the scale on the university students. Participants were asked to report on anticipated positive and negative affect around a hypothetical event (such as how their moods change over time). Accordingly, a parameterization of emotional vacillation items was performed by using the graded response model. Using the discrimination parameters and item fit statistics, some items were removed from the original scale, and a ten-item emotional vacillation version was developed. An analysis was carried out using the graded response model (GRM) in order to give a good fit for the scale and using flexMIRT to estimate the parameters. It was found that the GRM provided a better fit to the data. The reliability values computed based on the classical approach and IRT were above .80 after the item elimination process with only a minor drop. It was also found that none of the items showed DIF between males and females. It was concluded that the emotional vacillation scale was a valuable measurement tool to determine how the mood of university students changes and/or fluctuates often time, as this would also allow academicians and researchers to conduct research with this population using EVS.*

KEYWORDS: Item response theory, Graded response model, Emotional vacillation, Undergraduates.

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 3, 2025 (pp. 72-85)

www.abjournals.org

INTRODUCTION

Our emotions play a big role in our day-to-day existence. They add flavour to everyday interactions and enable us to respond appropriately to changes in our surroundings (Kuppens & Verduyn, 2015). As a result, our capacity for feeling emotions is essential to day-to-day functioning and plays a significant role in determining our psychological and mental health. The function that emotions play in individuals' daily lives is significant. They not only make ordinary events more interesting, but they also help us react correctly to environmental changes. As a result, feeling emotions is crucial for day-to-day functioning and is a major factor in defining our psychological and mental health. Emotional vacillation refers to the phenomenon of experiencing frequent and significant shifts in one's emotional state or mood, often characterized by sudden and extreme changes in emotions. It can involve rapidly transitioning from positive to negative emotions or vice versa or experiencing a continuous oscillation between various emotional states (Dåderman, et al, 2025). Emotional vacillation is a complex and multifaceted concept, and it can have various causes and implications for undergraduates' psychological well-being (Kuppens, Oravecz & Tuerlinckx, 2010).

Literature have shown that when undergraduates do experience mood swings on a regular basis, it tends to impact their relationships with teachers and fellow students, which may further impair their academic performance. Research aims to lessen students' tendency to waver in their emotions, which may also have a detrimental impact on their mental health. The literature's tendency demands the development and application of tools to gauge university students' emotional vacillation. For this purpose, Emotional vacillation scale (EVS) that was coined from the literature of Kuppens & Verduyn (2015) and Sullivan and Strongman (2003) which was used to measure frequent and significant shifts in one's emotional state or mood among high school students, and it has also been used in various studies.

Validity and reliability of our measuring instrument in the field of psychology are highly germane. Over ten decades now, the testing industry has benefited greatly from the application of classical test theory (CTT). However, the development of psychological tests has undergone significant and beneficial modifications as a result of the application of item response theory (IRT) to psychological and educational assessment. When model fit is present, IRT produces person parameter invariance, which means that test scores are independent of the specific choice of test items. Additionally, test information functions provide the amount of information or "measurement precision" captured by the test on the scale measuring the construct of interest, among other features (Embretson, 1996; Hambleton et al., 2000; Li, et al, 2024). These are the main advantages of using IRT with test development over CTT.

One important psychometric technique used in the processes to produce trustworthy assessment instruments is IRT. Test developers have found widespread usage for it because it addresses a variety of measurement concerns that come up during the test development process and yield a richer output (Embretson, 1996; Samejima, 1968). IRT was used in the validation of the EVS for undergraduates because of this. However, the accuracy of the measurement in IRT depends on the amount of the latent characteristic that is being measured. Numerous models have been developed to evaluate Likert-type items that feature more than two response categories, or polytomous items. These polytomous response models differ in their parametrization, but they all provide a location parameter and fixing the slope

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

to 1, (as well as the associated characteristic curve) for every response category (Thissen & Steinberg, 1986). For this study, the graded response model was used because GRM's basic concept is to apply a probabilistic function to each response category in order to extend the logic of more straightforward dichotomous IRT models to polytomous scales.

Assumptions of IRT

Unidimensional IRT applications are predicated on two fundamental tenets: unidimensionality and the form of the item characteristic curves (ICC). It is affirmed that a test's set of items measures only one latent construct. It is therefore expected that there would be a dominating factor that accounts for many of the instrument variance scores. Eigenvalue plots (20 percent or greater variability on the first component, parallel analysis, and confirmatory factor analysis (proving the hypothesis of a single factor) are some popular methods used to verify unidimensionality (Hambleton, et al, 1991; Hattie, 1985). The second supposition holds that the model-specified ICC represents the connections between the latent characteristics and the item responses. Examinee success rates (ICCs) typically resemble "S" curves and correspond to examinees' ability-based probability when test items are binary scored (e.g., true-false or yes-no). Easy test items are moved to the left of the trait's measuring scale, and difficult test items are moved to the right of the scale. Higher slopes are found in discriminating items compared to lower discriminating items. When the model fit is right, the ICC strongly tends to match up with the test data.

Graded Response model

A polytomous IRT model called the Graded Response Model (GRM) was developed for item replies represented by graded categories (Samejima, 1968). It is affirmed that the likelihood of providing a correct response is modeled as a function of θ, item difficulty (b), and item discrimination (a) in dichotomous two-parameter logistic (2PL) IRT models. Numerous studies have demonstrated that the model works better when applied to Likert-type items (Rubio et al., 2007; Mielenz et al., 2010). The reason why GRM was chosen for the current investigation despite the fact that there are a few other alternative polytomous models that are accessible in the literature is because GRM assumes unidimensionality, which means it is specifically designed for scales with polytomous, ordered responses (e.g., "strongly disagree," "disagree,", "strongly agree", agree). It estimates item difficulty and respondent severity, allowing for the examination of both item and person characteristics and the assessment of measurement precision for scales with ordered response categories. If the items in the instrument are designed to measure a single construct or trait just like the one that is measured in this study which is emotional vacillation, GRM is a suitable choice. The polytomous extension of the S-X^2 item-fit index could also be used to evaluate the item level fit statistics, which may also be derived using the IRT-based technique (Orlando & Thissen, 2000). This index uses a significance test and is based on the Chi-square method, with *p*-values less than.05. were typically interpreted as a sign of inadequate item fit.

Considering the aforementioned benefits of GRM, using IRT in test development and revision procedures offers advantages over classical test theory. Consequently, within the past ten years, there have been more studies that have used IRT for test development, test revision, and obtaining shorter versions of existing tests (Zanon, Hutz, Yoo & Hambleton, 2016; Istiyono et al., 2019; Bilker et al., 2012). Given this, the primary goal of this study is to

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

revise, validate and fit EVS utilizing an IRT-based methodology called Graded Response Model (GRM).

The 2PL model is used to construct the GRM in model fitting. This model, which is regarded as a generalization of the two-parameter logistic model (2PL), is suitable for usage when dealing with ordered categories on a rating scale (such as a Likert scale reflecting levels of agreement or disagreement) (Keller, 2013). The 2PL model is used to give the likelihood that an individual, given the degree of the underlying latent trait, will earn a particular score, or higher. The more of the trait possessed by respondents, the more likely they are to respond with answers that receive higher scores, or the more likely they are to choose one of the more positive ratings on the items of the scale.

In GRM, the probability P_{jb} (θ) that an individual's response falls at or above a particular ordered category given latent trait (θ) could be written in the equation below:

$$P_{jg}^{*}(\theta) = P(Y_j \geq g \mid \theta) = \frac{\exp[a_j(\theta - b_{jg})]}{1 + \exp[a_j(\theta - b_{jg})]}$$

Parameter a_j is the item discrimination parameter and $b_j g$, g = 1,…,$m\_\ j$, are often referred to as the threshold parameters. It should be noted that the parameter a_j is constant for all categories of the same item. Different items will perhaps present different discriminations. The $b_j g$ parameters are the latent traits in which the probability of answering at or above the particular category equals 50%.

METHOD

Study group

The population of the study comprised 561 undergraduates that were randomly selected in Nigerian universities. This comprised 206 males and 355 females whose age ranges between 16 years to 30 years. A cross-sectional research design was adopted, and a non-probability convenient sampling was used to select the students in their various departments. Because IRT investigations rely on models, a sizable sample size is crucial to the measurement's accuracy. A sample size of 561 is said to be enough for models with more estimated parameters. This was in line with Reise & Yu (1990) who suggested that a minimum of 500 examinees is appropriate for a GRM. Because of this, the study's sample size was purposefully kept large, and participation was entirely voluntary.

Measurement instrument

EVS was developed by the author of this study. The instrument comprised ten items with a four-point Likert response format ranging from strongly agree, agree, disagree and strongly disagree. The Cronbach alpha coefficient was estimated as .81. The item-elimination process followed a rigorous psychometric validation procedure using both classical and Item Response Theory (IRT) criteria. The initial item pool consisted of 15 items which were developed based on a thorough literature review and expert input to capture various dimensions of emotional vacillation among undergraduates. During preliminary analyses, items were screened for low item-total correlations, poor factor loadings, and weak

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

discrimination parameters (a < 0.50) in the Graded Response Model. Five items (Items 3, 7, 9, 11, and 14) were eliminated due to redundancy, low discrimination, or inconsistent threshold ordering that suggested unclear response functioning. These eliminated items are: I rarely experience changes in my mood during the day *(3), I only feel emotionally unstable when I am physically tired (7)*, Once I feel upset, I stay that way for a long time *(9)*, I can easily predict how I will feel in different situations *(11), and, My emotional instability is mostly due to external influences (14)*. The remaining items demonstrated strong psychometric properties, including acceptable item-fit indices, ordered category thresholds, and satisfactory internal consistency. Following this trimming, the final Emotional Vacillation Scale retained 10 items, each contributing meaningfully to the underlying construct.

To strengthen evidence for construct validity of the 10 items, content validation involves ensuring that the items on the *Emotional Vacillation Scale* adequately represent the theoretical construct of emotional instability and fluctuation relevant to undergraduates in Nigeria. Two experts in the field of psychology, educational measurement, and cultural studies reviewed the items for clarity, relevance, and cultural appropriateness, confirming alignment. Feedback from these experts provided evidence that the scale's content validly captures the intended emotional construct before applying Item Response Theory analyses.

Procedures

The participants were asked to respond to the EVS items, beginning with questions about their demographics. The students were made aware that their involvement in the study was entirely voluntary and that the information they submitted would be kept private. Undergraduates were instructed to carefully read the questions and complete the scale in accordance with their opinions. For every class, data collection was placed in a single session. For the ICRF and ICC, the ability levels were set from -3 to 3.

Analysis

To test the unidimensionality of the data, a confirmatory factor analysis was performed using R. The fit of the model was evaluated using the $\chi 2$ statistic, Root Mean Square Error of Approximation (RMSEA), standardized root mean squared residual (SRMR) and confirmatory fit index (CFI). In addition, unidimensionality was also evaluated by applying an exploratory factor analysis. The GRM was used for IRT, while for the item calibration flexMIRT was used. For the Differential Item Functioning, the Mantel-Haenszel test was used. After obtaining item parameters from flexMIRT, the GRM formula stated below was used to calculate the probability of each of the item.

$$P_{jg}^{*}(\theta) = P(Y_j \geq g \mid \theta) = \frac{\exp[a_j(\theta - b_{jg})]}{1 + \exp[a_j(\theta - b_{jg})]}$$

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 3, 2025 (pp. 72-85)

www.abjournals.org

RESULTS

Fitting the graded response model

The EVS has four graded response options from 0-3. Hence, the GRM was preferred for model fitting because it helps researchers validate the scale by providing insights into how well each item is measuring the underlying construct and identifying problematic items or response patterns. As a result of the GRM estimations, one discrimination and three threshold parameters were obtained for each item. The result below shows the graded item of EVS scale. While h2 is communality, "a" is the discrimination parameters while d1, d2 and d3 are the beta. It could be interpreted that the beta is increasing the GRM forces the threshold to be ordered. The analysis showed that the fit values of the single-factor GRM model were at an acceptable level: CFI =0.942 TLI =0.954, RMSEA=.026, and SRMR =0.067. The Marginal reliability for response pattern scores: 0.59, which measurement is somewhat reliable.

The communality values ranged from .356 to .704. The results provided evidence that the items are fairly different in terms of the amount of variation with a common factor. The discrimination parameters ranged from 0.19 to 1.25. The contribution of items to total information varied between 2.08 and 2.45. This finding shows that the contribution of items to the model shows a significant variance. The difficulty parameters of the emotional vacillation scale vary between -1.31 and 13.78. This finding shows that the emotional vacillation scale is useful to some extent for identifying individuals with unstable emotion but was a scale more suitable for identifying average-to-high-level individuals who often do not have stable emotion.

Table 1: FlexMIRT result

Item		flexMIRT item parameters			
	h2	a	d1	d2	d3
1	.432	1.25	-1.71	-0.01	2.15
2	.356	0.78	-2.90	-0.52	1.58
3	.704	0.19	-2.60	5.04	13.78
4	.638	0.76	-3.21	-0.61	1.71
5	.603	0.56	-2.28	1.03	4.69
6	.456	0.47	-1.73	0.57	4.61
7	.460	0.29	-4.21	1.16	9.83
8	.624	0.50	-5.56	-1.31	2.54
9	.396	0.97	-1.70	0.09	2.18
10	.707	0.50	-2.79	-0.37	3.53

h2= communality

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 3, 2025 (pp. 72-85)

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.83
Bartlett's Test of Sphericity	Approx. Chi-Square	2155.579
	Df	45
	Sig.	<.001

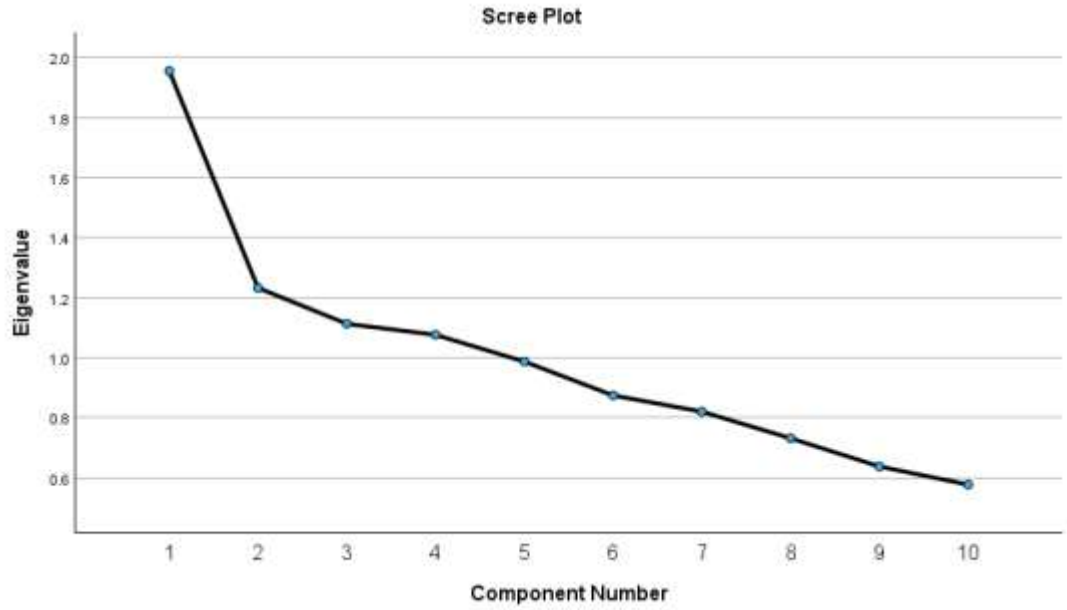


Figure1: *A scree plot indicating a gradual decline in eigenvalues as more factors are considered.*

Testing IRT assumptions

The confirmatory factor analysis (CFA) results showed that a one-dimensional model was confirmed at acceptable level: χ 2 (N=561, df=45) =2155.579, p<0.01. Factor loadings varied between 0.36 and 0.70 at p<0.001 significance level. In addition, an exploratory factor analysis (EFA) was conducted as a supplemental. The results showed that two factors were extracted with eigenvalues greater than one. At this point, IRT based analyses were performed to obtain more acceptable results to keep original one-dimensional structure of the EVS for university students. This was achieved by investigating the information contribution of each item and item level fit Statistics.

Item selection for EVS for university students

The items were selected from the 10 item EVS to increase the model fit index, contributing to the validation process of the scale for university students. Two different criteria were used in this process. First, the amount total item information was obtained by adding of the information values of each item. The remaining ten items were added back to the GRM model, item fit statistics were looked over, and any items with a low fit level were identified and removed. Upon repeating the GRM model with the remaining ten items, it was discovered that none of the items needed to be deleted based on this criterion.

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 3, 2025 (pp. 72-85)

ICRF of four response categories of an item

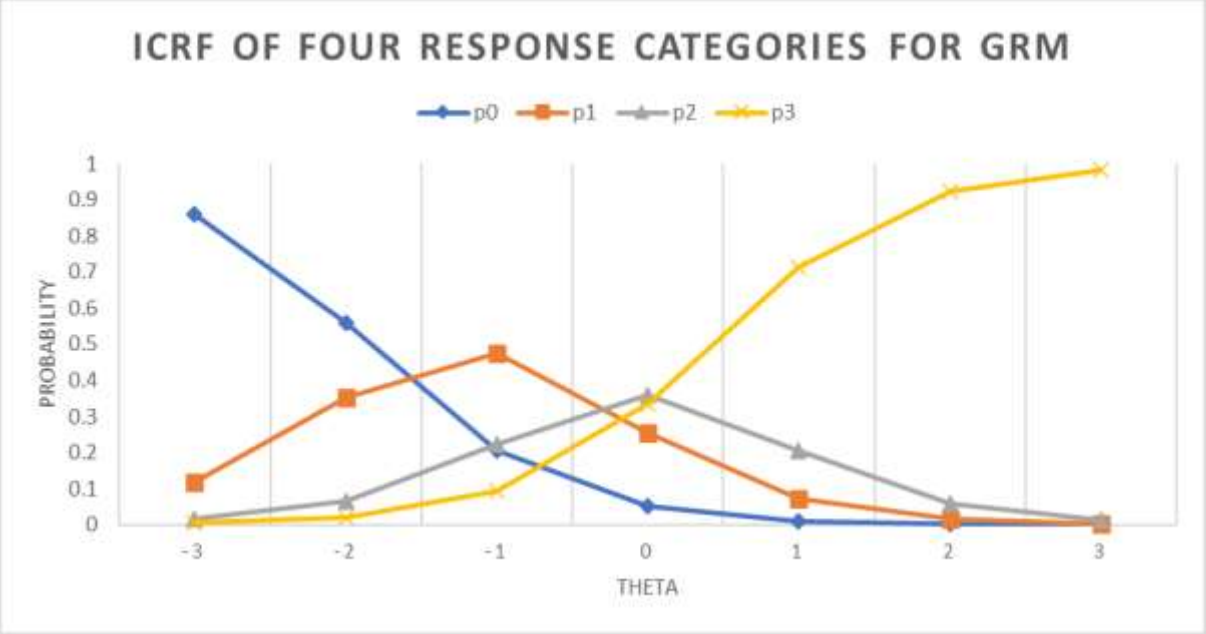


Figure 2: *Category response curves for a four-category graded response model (GRM) for item 2 ("When I feel more positive emotion (such as joy or amusement), I change what I am thinking about"), with a = .78, d1 = -2.90, d2 = -0.52, and d3 = 1.58*

Figure 2 reveals a graphical representation of the item category response functions (ICRF) (in terms of probabilities of each rating category over the latent trait scale) for an item and shows an example of how the researcher will evaluate negative and positive emotion of items. The latent trait is typically presented in the standardized form that goes from -3 to 3. The mean of trait scores on the construct of interest is set to 0 and the standard deviation to 1. These scores can always be transformed later to a more convenient scale, and I think they should be similar to many psychological scales. The "a" parameter for this item is .078 and the d parameters or thresholds for the first, second, and third categories are respectively -2.90, -0.52, and 1.58. Again, the first category is not estimated since the probability of getting a score of zero or higher is always 1.0. CRF represents the probabilities for responding to each of the five response categories as a function of respondents' level of latent trait. To cite an example, a person with high levels of emotional vacillation is more likely to choose higher values (as, it was revealed in the item) on the Likert type item that: *When I feel more positive emotion (such as joy or amusement), I change what I am thinking about*." It should be noted that CRF could be used to identify items with low category discrimination and spread.

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 3, 2025 (pp. 72-85)

Item Characteristic Curve for ten items

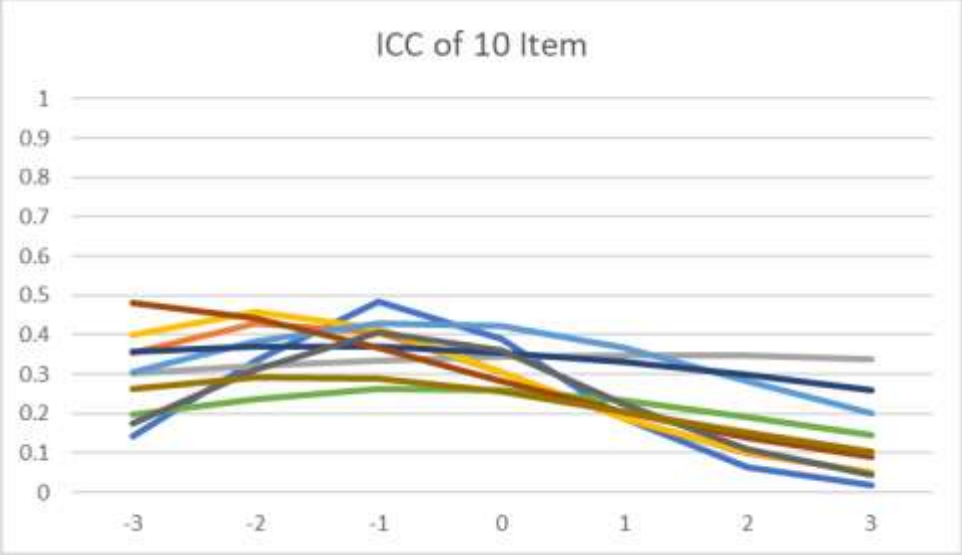


Figure 3: Showing the ICC for the ten Items

Figure 3 revealed the general downward trend of most lines from left (−3) to right (+3) suggests that as respondents' latent trait increases, their probability of choosing lower response categories decreases. The Item Characteristic Curves (ICCs) for the 10 items of the EVS indicate that the items demonstrated moderate discrimination across the latent continuum. As the latent trait level increased, the probability of endorsing lower response categories decreased consistently across items.

Test Information Curve

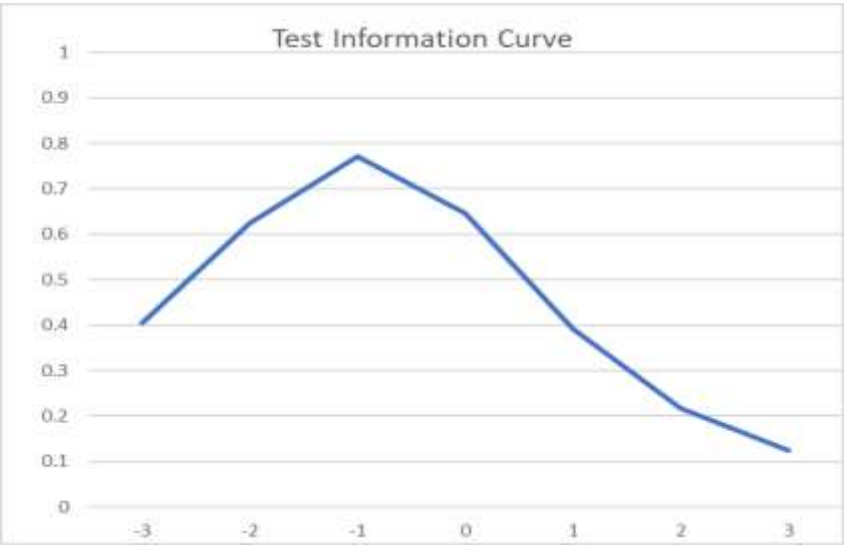


Figure 4: Item Information Curve

As revealed in Fig. 4, most of the information provided by the scale was below the mean of respondent scores indicating that the scale was better designed for respondents with lower scores. The Test Information Curve (TIC) for the EVS reveals that measurement precision

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 8, Issue 3, 2025 (pp. 72-85)

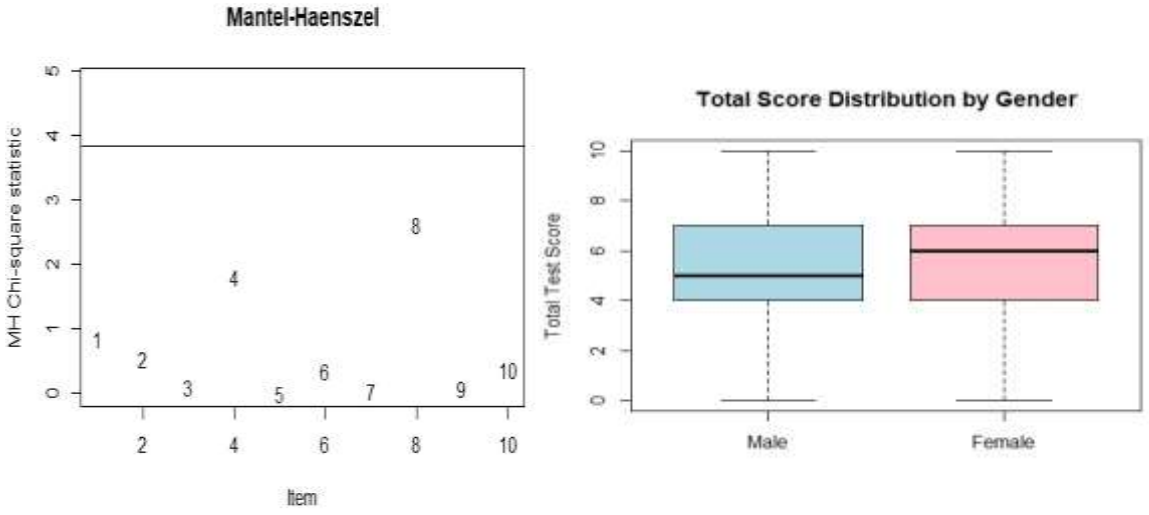
was highest around θ = -1, which implies that the instrument provides the greatest information for respondents exhibiting slightly below-average levels of emotional vacillation.

Mantel-Haenszel Chi-square statistic

Table 2: DIF analysis

Items	*Stat.*	*P-value*	*Remark*
1	0.8389	0.3597	No DIF
2	0.5324	0.4656	”
3	0.0920	0.7616	”
4	1.8001	0.1797	”
5	0.0003	0.9873	”
6	0.3416	0.5589	”
7	0.0334	0.8550	”
8	2.6092	0.1062	”
9	0.0718	0.7887	”
10	0.3608	0.5481	”

Figures 5 & 6



The results from Table 2 and figure 5 reveal the output of the MH^2 values for the whole 10 items. It could be revealed that none of the items was detected as DIF. This means that the test items function equivalently across groups. Test takers with the same ability level regardless of their gender (male and female) have an equal probability of answering the item correctly (or endorsing a response category). Figure 6 revealed that the Emotional Vacillation Scale performs equally well for both male and female undergraduates. The items measure the same underlying construct (emotional vacillation) without bias. The differences in total scores between males (5.41) and females (5.49) can be interpreted as due to true differences in the underlying trait (emotional vacillation), not bias in the item wording or functioning.

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

Each item is fair across groups. Gender does not influence how the item behaves after controlling for overall ability.

DISCUSSION

The main goal of this study was to validate the 10-item Emotional Vacillation Scale for university students, which was modified from Sullivan and Strongman's 2003 literature on mixed emotion. Item response theory (IRT) was applied in this sense to derive conclusions about the scale's psychometric qualities. Certain components that interfered with the one-dimensional structure and produced information that was below average were eliminated using IRT-based procedures. The amount of information provided by the scale and its psychometric qualities were unaffected by the item removal process based on IRT. Furthermore, it was anticipated that the scale's reduction made it more beneficial. Shorter scales are more beneficial in practice because they allow for the inclusion of more variables in the study and broaden the nomological network of emotions to include additional psychological constructs using scales specifically designed to measure those constructs. Ultimately, this validation study enabled researchers who are interested in doing so to measure university students' emotional vacillation. There isn't a standardized emotional vacillation scale that is appropriate for university students, despite the fact that there are emotion and mixed emotion measures that can be utilized with them. This research will help close this gap in the body of knowledge.

The discrimination parameters obtained from the IRT analyzes showed that the ten-item emotional vacillation scale validated for the university students was effective in distinguishing students whose emotion changes often due to academic work or other personal issues. On the other hand, when the IRT-based difficulty parameters were examined, it was seen that the scale provided more accurate measurements for average-to-high-level individuals. This means that the lower the scores, the better the scale for the students. In the current study, Differential Item Functioning (DIF) was investigated. DIF occurs when different subgroups of participants (male and female) with the same latent trait level yield different response patterns. However, it was revealed that there was no DIF between male and female undergraduates. It could be inferred that the scale demonstrates measurement invariance and gender fairness, supporting its validity for use among Nigerian undergraduates.

This study contributed to literature and experts working in the fields of psychology of emotion and psychometrics as it provides myriads of information regarding the students as well as the scale in many different ways. Firstly, as emphasized by different experts, understanding the importance of the concept of emotional vacillation and its relationship to other psychological structures is vital as it goes a long way in determining how students perform in their academic works. Even though there are studies forming the nomological network of what causes sudden changes in students' emotions with other psychological constructs, there is an absence of literature on the construction of emotional vacillation. In addition, there are appropriate instruments for measuring emotion among high school students, while there is no instrument for such studies to be carried out with university students. This validated instrument has contributed to a better understanding of emotional vacillation of university students and also understanding of the relationship between

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

emotional vacillation and other factors that could either impinge or enhance the academic work of undergraduates.

The scale's usefulness rose with fewer elements while the majority of the information it contained remained same. Furthermore, despite the removal of some items, the scale's reliability level remained rather constant. It makes sense that the researcher and practitioners would prefer a scale that is shorter but yet valid. That being said, educationist and other stakeholders who plan to use or adopt this scale in their research should be aware of the limitation of this study. The data used in this analysis was provided only by students enrolled in public universities located in South-West, Nigeria. Therefore, this has a limit on the generalizability of this study. However, as Baker (2001) pointed out, IRT parameters are unaffected by the sample data that was gathered. As a result, the parameters acquired are unaffected by the sample size. Therefore, given that the IRT was used, it may be concluded that the results' generalizability may not pose a significant issue.

CONCLUSION

For items with ordered response categories, the Graded Response Model works well and has made it possible to analyze the emotional vacillation construct in detail. The model provided a more accurate assessment of emotional vacillation among undergraduates by accounting for both the degree of endorsement and the participants' inclination to endorse an item. The outcomes of the IRT validation process include a number of significant findings. First, the analysis revealed that the EVS is a suitable tool for assessing the intended construct in undergraduate Nigerian students. The scale's validity in this cultural and educational setting is supported by the items' discrimination and difficulty parameters, which match the underlying trait. Moreover, components that contribute somewhat or substantially to the overall measurement precision have been identified through the application of IRT. In addition, it revealed that there was no DIF in the items as well as between male and female participants which means that there was measurement invariant.

In conclusion, the Emotional Vacillation Scale (EVS) for Undergraduates in Nigeria has been validated using the Graded Response Model in the context of Item Response Theory. The nuanced insights gained from this analysis not only affirm the appropriateness of the scale for measuring emotional vacillation but also offer guidance for potential refinements, ensuring its continued relevance and effectiveness in assessing the targeted construct within the unique cultural and educational context of undergraduates in Nigeria, USA and other countries of the world. Future research should therefore include more diverse samples across multiple geopolitical zones to enhance representativeness and strengthen the generalizability of the validated instrument.

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

REFERENCES

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae.net/irt/baker..

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354-369.

Dåderman, A. M., Persson, B. N., Ahlstrand, I., Hallgren, J., Larsson, I., Larsson, M.,... & Pennbrant, S. (2025). Item response theory modelling of the trait emotional intelligence questionnaire-short form: item streamlining, differential item functioning, and validity in a Swedish multicenter cross-sectional study. *BMC psychology*, *13*(1), 987.

Embretson, S., & Reise, S.P. (2000). Item Response Theory for Psychologists. Lawrence Erlbaum Associates. Inc. Mahwah.

Embretson S. E. (1996). The new rules of measurement. *Psychol Assess*. 1996;8 (4):341–9. doi:10.1037/1040-3590.8.4.341.

Gruber, J., Kogan, A., Quoidbach, J., & Mauss, I. B. (2013). Happiness is best kept stable: Positive emotion variability is associated with poorer psychological health. Emotion**,** 13(1), 1-6.

Hambleton RK, Swaminathan H, & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Thousand Oaks: Sage Publications.

Hanson, T. A. (2024). Interpreting and psychometrics. In *The Routledge Handbook of Interpreting and Cognition* (pp. 151-169). Routledge.

Hattie J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Appl Psychol Meas*.; 9:139–64.

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin***,** 141(4), 901-930.

Keller, L. A., & Hambleton, R. K. (2013). The Long-Term Sustainability of IRT Scaling Methods in Mixed-Format Tests. *Journal of Educational Measurement*, *50*(4), 390-407.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology,* 99(6), 1042-1060.

Kuppens P., Verduyn P. (2015). Looking at emotion regulation through the window of emotion dynamics. *Psychological Inquiry*, 26, 72–79.

Li, Y. Y., Tong, L. K., Au, M. L., Ng, W. I., Wang, S. C., Liu, Y., ... & Qiu, X. (2024). Psychometric evaluation of the study interest questionnaire-short form among Chinese nursing students based on classical test theory and item response theory. *BMC nursing*, *23*(1), 717.

Mielenz, T.J., Edwards, M.C. & Callahan, L.F. (2010). Item response theory analysis of two questionnaire measures of arthritis-related self-efficacy beliefs from community based US samples. Hindawi Publishing Corporation Arthritis.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied psychological measurement*, *24*(1), 50-64.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of educational Measurement*, *27*(2), 133-144.

Rubio, V.J., Aguado, D., Hontangas, P.M., & Hernandez, J.M. (2007). Psychometric properties of an emotional adjustment measure. *European Journal of Psychological Assessment*, 23 (1), 39-46.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores 1. *ETS Research Bulletin Series*, *1968*(1), i-169.

Samejima, F. (1969) Estimation of Latent Ability Using a Response Pattern of Graded Scores. (Psychometrika Monograph, No. 17). Psychometric Society, Richmond. http://www.psychometrika.org/journal/online/MN17.pdf.

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148

Sullivan, G. B., & Strongman, K. T. (2003). Vacillating and mixed emotions: a conceptual-discursive perspective on contemporary emotion and cognitive appraisal theories through examples of pride. *Journal for the Theory of Social Behaviour*, *33*(2), 203-226.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51(4), 567–577.

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(0), 18.

Istiyono, E., Dwandaru, W. S. B., Lede, Y. A., Rahayu, F., & Nadapdap, A. (2019). Developing IRT-Based Physics Critical Thinking Skill Test: A CAT to Answer 21st Century Challenge. *International Journal of instruction*, *12*(4), 267-280.

EMOTIONAL VACILLATION SCALE (EVS)

This is not a test, it has no right or wrong answer. The scale is examined to know or determine the emotional waver, fluctuation or change of the students. **Note SA = Strongly Agree, A = Agree, D = Disagree, SD = Strongly Disagree.**

S/N		SA	A	D	SD
1	My emotion changes often time which I do not have control over				
2.	When I feel more positive emotions (such as joy or amusement), I change what I am thinking about.				
3.	I feel different emotions occur very quickly one after another				
4	When I feel less negative emotion (such as sadness or guilt), I change what I am thinking about.				
5	I have difficulty predicting how I will feel from one day to the next.				
6	I find it challenging to maintain a consistent emotional state over time.				
7	I control my emotions by changing the way I think about the situation I am in.				
8	When I am feeling negative emotions, I make sure not to express them so as not to affect my academics.				
9	When I want to feel less negative emotion, I change the way I think about the situation.				
10	I often find myself feeling differently than I did just a few hours apart.				

Article DOI: 10.52589/BJELDP-6LKXN148
DOI URL: https://doi.org/10.52589/BJELDP-6LKXN148