www.ijses.net

Effect of Anchoring Vignettes on Reliability and Validity as a Measure of Motivation and Emotion in Mathematics

Khajidmaa Otgonbaatar
Mongolian National Institute for Educational Research, Mongolia

Itgel Miyejav
Mongolian National Institute for Educational Research, Mongolia

To cite this article:

Otgonbaatar, K. & Miyejav, I. (2024). Effect of anchoring vignettes on reliability and validity as a measure of motivation and emotion in mathematics. *International Journal of Studies in Education and Science (IJSES), 5*(4), 477-489. https://doi.org/10.46328/ijses.113

IJSES International Journal of Studies in Education and Science
2024, Vol. 5, No. 4, 477-489 https://doi.org/10.46328/ijses.113

Effect of Anchoring Vignettes on Reliability and Validity as a Measure of Motivation and Emotion in Mathematics

Khajidmaa Otgonbaatar, Itgel Miyejav

Article Info

Article History

Received:
15 February 2024
Accepted:
24 June 2024

Keywords

Anchoring vignette
Composite reliability
Convergent validity
Discriminant validity

Abstract

In mathematics, the importance of emotion and motivation has long been appreciated. Meanwhile, the measurement of these constructs needs to be carefully developed. Indeed, many previous studies employed questionnaires and self-rating to investigate emotion and motivation in mathematics. However, researchers stressed that self-rating and questionnaires are vulnerable to different biases, such as cultural bias, social desirability bias, and individual reporting behavior, which could influence the reliability and validity of the responses. The present study employed an anchoring vignette approach and self-rating to measure emotion and motivation to examine its effect on reliability and validity. The study sample consisted of 308 Grade 8 students from urban, suburban, and rural schools in Mongolia. The students were administered two vignette sets and their respective self-rating items in 45 minutes. The study found that the anchoring vignette approach increased two types of reliability coefficients, including composite reliability and Cronbach's α, and two types of validity, such as convergent and discriminant validity. In sum, the results of the present study concluded that the anchoring vignette approach has the potential to improve the reliability and validity of tools to measure emotion and motivation in mathematics.

Introduction

There has been growing interest in emotion and motivation in mathematics in the last decade (Schukajlow et al., 2017). Emotions and motivation are conceptually overlapped; however, these two constructs are different phenomena (Schukajlow et al., 2017). According to Schukajlow et al. (2017), "emotion theories consider emotion-specific motivational impulses as part of emotion, and motivation theories consider emotions as psychological forces that motivate action" (p. 4). Enjoyment has been one of the most stated positive emotions in the learning context (Pekrun et al., 2002). Mathematics enjoyment is "… the positive emotions or pleasure a student experience while solving mathematical problems" (Otgonbaatar, 2021a, p. 69). Specifically, enjoyment is a positive emotion and is regarded as an essential part of interest (Krapp, 2005). Therefore, enjoyment can be a predictor of an individual's interest in mathematics (Schukajlow & Rakoczy, 2016). Emotional responses to mathematics can cover positive emotions, such as enjoyment and excitement, and negative emotions, such as panic and boredom (Grootenboer & Marshman, 2016). Moreover, emotions are an influential part of learning mathematics, and consistent emotional responses to mathematics result in a more stable attitude toward

mathematics (Grootenboer & Marshman, 2016). Particularly, Cheeseman & Mornane (2014) found that the more the students have positive emotions, the greater they persist in mathematical learning. According to Pekrun et al. (2017), math enjoyment positively affected students' grades and test scores in mathematics. If students experience a lack of enjoyment in mathematics class, it may lead to mathematical avoidance and mathematical anxiety (Wadlington & Wadlington, 2008).

From the motivational framework, self-efficacy is one of the most frequently examined motivational constructs in education (Schukajlow et al., 2017). On the contrary, it has been disregarded in mathematics education for many years (Zan et al., 2006). Mathematics self-efficacy is "… the belief of a person in his/her competence to solve mathematical problems and tasks successfully" (Zimmermann, Bescherer, & Spannagel, 2014, p. 3). There is a large body of literature indicating that students who feel competent in math tend to perform better in math (Kitsantas et al., 2010; Skaalvik, Federici, & Klassen, 2015; Yıldırım, 2012). On the other hand, student's math achievement influences their math self-efficacy. Following the discussion above, the present study focuses on math enjoyment and self-efficacy as essential emotional and motivational constructs in mathematics.

Meanwhile, measuring these constructs needs to be carefully developed (McLeod, 1994; Otgonbaatar, 2020). In mathematics education, McLeod (1994) pointed out from the psychometric perspective that "… questionnaire data were not necessarily reflecting accurately what students were thinking and feeling" (p. 640). Certainly, a large body of previous studies employed questionnaires and self-rating as a tool to investigate math self-efficacy and math enjoyment. However, researchers have stressed that self-rating and questionnaires are vulnerable to different biases, such as cultural bias, social desirability bias, and individual reporting behavior, which could influence the reliability and the validity of the responses (see, e.g., Paulhus, 1991; Primi et al., 2016; Weiss & Roberts, 2018). Respondents might give fake answers to the questionnaires. For instance, when students are requested to rate themselves on items such as "I am confident to solve a geometry problem in my math class" or "I enjoy solving math problems like finding the area of a triangle," they might opt for the highest category of the scale to attract teachers or observers (West et al., 2016). From the cultural perspective, researchers found that East Asians, specifically students from Japan and South Korea, are likely to choose the midpoint of the Likert-type scale (Chen et al., 1995). The biases affect the actual level of an examinee's response and lead to poor quality of the scores (West et al., 2016).

To tackle the above challenges related to measurement, researchers introduced a novel approach called "anchoring vignette" [AV] in social sciences (King et al., 2004). AV is a short description or scenario about fictitious individuals who hold different levels of skills or facets. The respondents are asked to evaluate the individuals characterized in the vignettes applying the same scale for self-rating items. Then, the respondents' self-ratings are transformed compared to how they rated the individuals in the vignettes using a specific scoring rule explained in the method section. Researchers (Vonkova & Hrabak, 2015; Otgonbaatar, 2021b) noted that the application of AV approach is still rare in education, specifically in mathematics education, except PISA 2012. However, since AV approach was introduced, it has been widely used in other areas such as personality (Primi et al., 2016; Weiss & Roberts, 2018), job satisfaction (Kristensen & Johansson, 2008), health (Grol-Prokopczyk, 2014; Hinz et al., 2016; Pacheco, 2019; Poksinska & Cronemyr, 2017). Moreover, although previous studies found that AV

approach had a positive effect on reliability (Primi et al., 2016; Weiss & Roberts, 2018; Otgonbaatar, 2021b), the researchers encouraged to carry out more studies on its reliability and validity. The present study aims to investigate the reliability and validity of the AV approach in measuring math self-efficacy and math enjoyment.

Method

Participants

Participants were 308 Grade 8 students from 8 public schools in urban, suburban, and rural areas in Mongolia. Based on the background information, 49% of the participants were male and 51% were female. Participants were a mean age of 14 (SD=0.51, range=13-16).

Measures and Procedures

The present study employed three self-rating items for each scale such as math self-efficacy (e.g., "I feel confident in finding the area of a parallelogram.") and math enjoyment (e.g., "I think math is an enjoyable and cool subject.") followed by two vignette sets. Each vignette set consists of three vignettes, including low, medium, and high vignettes. The self-rating items for math self-efficacy were adopted from PISA-2003 items in Lee & Stankov (2013), and the items for math enjoyment were adopted from Grootenboer & Marshman (2015) and Kusmaryono et al. (2018). The vignette set for math self-efficacy was developed based on the guide for constructing self-efficacy scales developed by Bandura (2006). The vignette set for math enjoyment was developed on existing literature, which discussed characteristics of emotion and motivation in mathematics education (e.g., Krapp, 2005; Pekrun, 2006; Schukajlow et al., 2017;). The vignette sets for each scale is presented in Table 1.

Table 1. Vignette Sets for Math Self-Efficacy and Math Enjoyment

Construct	Vignettes
Math self-efficacy	**Low vignette:** Zaya feels he can't do it at all when he is assigned to solve problems like finding the size of angle x in a given figure. Generally, she often thinks that "I can't do it" when assigned to solve geometry problems. Based on this information, how much do you agree with the statement "Zaya is confident in solving geometry problems"? **Medium vignette:** Ganaa feels moderately confident that he can solve geometry problems like finding a triangle area. Because he feels somewhat unsure if he can proceed with some of the procedures to find the answer. Based on this information, how much do you agree with the statement "Ganaa is confident in solving geometry problems"? **High vignette:** Delger feels highly confident that she can accomplish the task when assigned to solve geometry problems like finding the height of a pyramid. Most of the time, she finds geometry problems too easy and unchallenging. Based on this information, how much do you agree with the statement "Delger is confident in solving geometry problems"?

Construct	Vignettes
Math enjoyment	**Low vignette:** Zoloo feels math is a boring and dull subject. He doesn't enjoy math class. When he hears the word mathematics, he has a feeling of dislike. He would like to spend less time in school doing mathematics. Based on this information, how much do you agree with the statement "Zoloo enjoys doing math"?
	Medium vignette: Sometimes, Gerel feels what she learns in math class is uninteresting and is mostly about numbers. So, she thinks math is not very enjoyable. In math class, she learns about the things that interest her. Based on this information, how much do you agree with the statement "Gerel enjoys doing math"?
	High vignette: Tsogoo looks forward to his math class. Because he is interested in the things, he learns in mathematics. He enjoys attempting to solve math problems, no matter if it is hard or easy. He also enjoys talking to other people about mathematics. Based on this information, how much do you agree with the statement "Tsogoo enjoys doing math"?

The participants were asked to rate themselves on the self-ratings items, and the hypothetical individuals in the vignettes on a 5-point Likert scale where 1=strongly disagree and 5=strongly agree. The participants completed the measures during the regular classroom activities. The data collection procedure was carried out based on the student's willingness with the permission of the school principals and math teachers.

Data Analysis

The current study utilized a simple non-parametric approach introduced by King & Wand (2007). The 5-point self-rating scales were converted into the vignette adjusted to a 7-point scale, as presented in Table 2.

Table 2. Possible values to self-rated and AV- adjusted item responses.

Responses to self-rating	Strongly disagree	Disagree	Neutral	Agree	Strongly agree		
	1	2	3	4	5		
AV adjusted responses	Lower than the low vignette	Same as low vignette	In between low and medium vignette	Same as the medium vignette	In between low and medium vignette	Same as high vignette	Higher than the high vignette
	1	2	3	4	5	6	7

Adopted from Kyllonen & Bertling (2013)

It should be noted that the above rule is applicable exclusively when the vignettes are evaluated in the intended

order (i.e., low vignette > medium vignette > high vignette). There might be ties or violations in the responses to the vignettes (for example, low vignette = medium vignette > high vignette) or vice versa (for example, medium vignette low vignette high vignette).

To address this issue, Kyllonen and Bertling (2014) proposed converting the old scale into the *AV*-adjusted new scale by selecting the lowest score among the range of potential scores. For instance, if a low and a medium vignette are tied, the range of attainable scores will be 2, 3, and 4. Then, the *AV*-adjusted new score is 2.

When order violation occurs in the vignette responses, for example, a low vignette is evaluated as higher than the medium vignette, or the medium vignette is evaluated as greater than a higher vignette. In such cases, the two vignettes are considered equal. However, it should be noted that the value assigned to the higher vignette should be applied to create the tie. The ties are then examined, as previously stated. The same technique was employed for this study. The "anchor" package for the R program, version 3.0-8, was used to analyze the data (Wand, King, & Lau, 2016).

Results & Discussion

Findings of Vignette Analysis

The vignette sets were first examined to test if the respondents comprehended the vignettes in the same manner as it was intended. This assumption is regarded as vignette equivalence, which is fundamental for further examination (King & Wand, 2007).

Table 3. Descriptive Statistics of Vignette Ratings for Math Self-Efficacy and Math Enjoyment

Vignette	Math self-efficacy			Math enjoyment		
	Mean	SD	n	Mean	SD	n
High vignette	4.69	0.76	308	4.76	0.60	308
Medium vignette	2.43	1.07	308	2.53	1.08	308
Low vignette	1.49	0.87	308	1.41	0.82	308

The means and standard deviations for self-rating and each of the three vignettes are shown in Table 3. The vignette orders and description adhere to consistency as shown by the vignette means. In other words, the assumption of vignette equivalence is supported since, on average, the high vignette is rated higher than the medium vignette, which is ranked higher than the low vignette.

Next, the vignettes' orderings for both constructs were examined. In Figure 1, the first bar presents "1,2,3" as the most common ordering, as 184 respondents (60%) for math self-efficacy rated the vignettes as intended. In the second bar, "{1,2},3" was the second most common ordering, with 86 respondents (28%) tying vignettes 1 and 2. Violations in vignette orderings for math self-efficacy are presented in bars 3, 5, 6, 7, 9, and 10. However, the violation in the ordering for math self-efficacy presented in less than 10% of the sample (9.4%). Order violation in the vignettes less than 10% is acceptable since it is considered a measurement error (Weiss & Roberts, 2018).

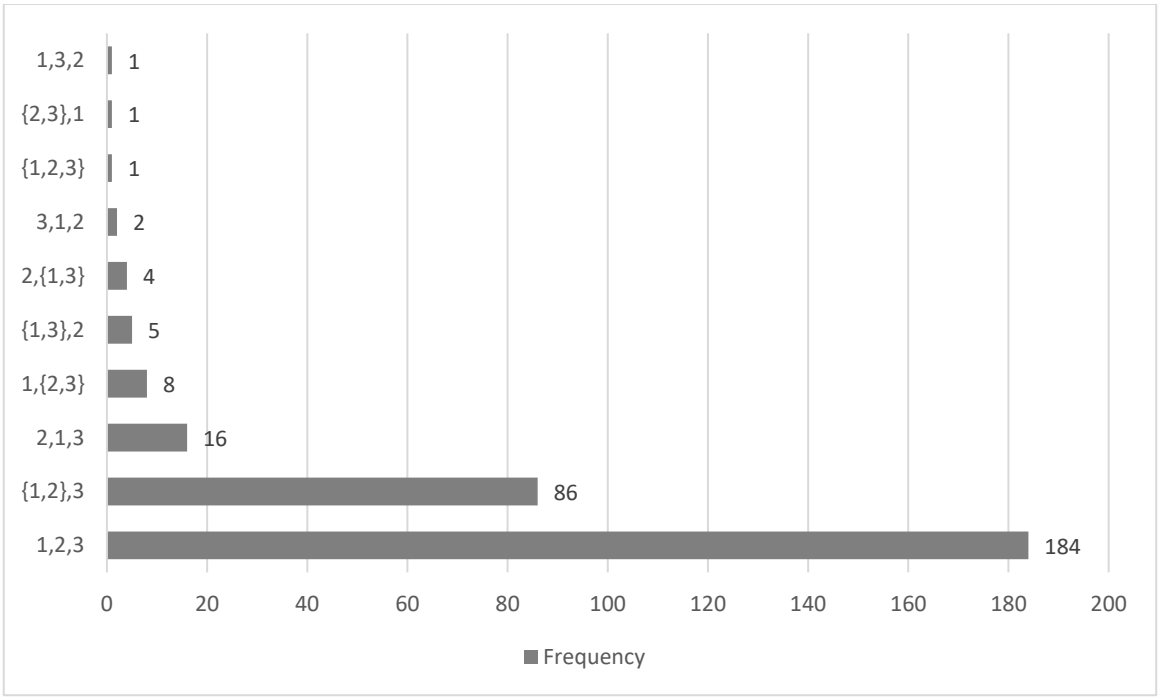

Figure 1. Vignette Orderings for Math Self-Efficacy (n=308)

Figure 2 illustrates the vignette ordering for math enjoyment, with the first row demonstrating '1,2,3' as the most frequent ordering as rated by 207 respondents (67%). The second most common ordering, as depicted by the second row, is '{1,2},3', with 65 respondents (21%) tying vignettes 1 and 2. The violations in vignette ordering occur in bars 3, 5, 6, 7, 9, and 10; however, they occur in less than 10% of the sample (8.1%). Overall, for both constructs, the vignettes functioned well and are applicable for further analysis.

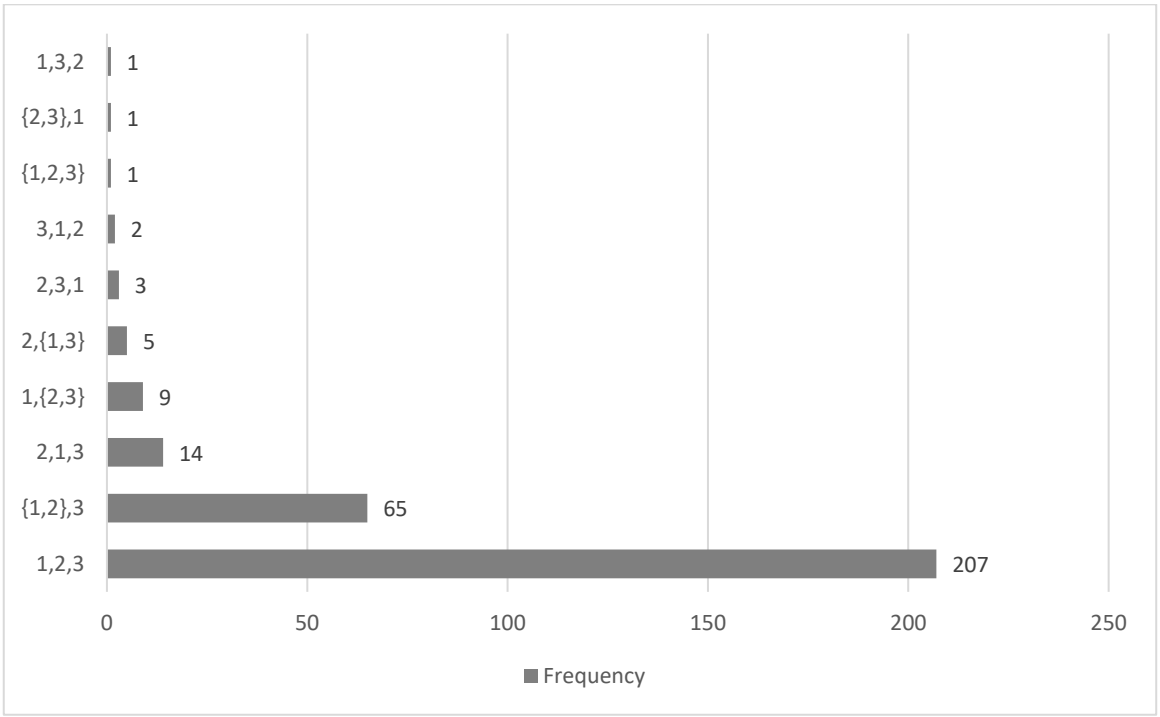

Figure 2. Vignette Orderings for Math Enjoyment (n=308)

Findings of Reliability and Validity Analysis

Reliability is an essential element of the validity of a survey instrument, which is regarded as the power of the instrument to measure a construct consistently (DeVon et al., 2007). Whereas validity is defined as "… the ability of the instrument to measure what it is supposed to measure for a construct" under study (Afthanorhan et al., 2014, p. 5). There are different types of validity in measurement. In this study, convergent and discriminant validity were examined since these two are suggested to be tested in measurement practice (Afthanorhan et al., 2014).

Convergent Validity

Convergent validity is "… correspondence or convergence between constructs that are theoretically similar" (DeVon et al., 2007, p. 5). To test the effect of the AV approach on convergent validity, the factor loading of two scales for each construct was analyzed first to determine convergent validity. For greater convergent validity, Hair et al. (2010) suggests that the factor loading of each item be at least 0.6. Hence, Exploratory Factor Analysis (EFA) was conducted for the original and AV-adjusted scale to determine factor loading math self-efficacy and math enjoyment. Factor loading for two scales shows two different constructs that are categorized into math self-efficacy (1st factor) and math enjoyment (2nd factor) without cross-loading (Table 4).

Table 4. Factor Loading for Original and AV-Adjusted Scale

Item	Original scale		*AV*-adjusted scale	
	1st factor Math self-efficacy	2nd factor Math enjoyment	1st factor Math self-efficacy	2nd factor Math enjoyment
1	0.689		0.851	
2	0.485		0.666	
3	0.629		0.747	
4		0.701		0.838
5		0.962		0.952
6		0.656		0.777

As presented in Table 4, after *AV* adjustment, factor loading for math self-efficacy ranged from 0.666 to 0.851, while for math enjoyment, the values ranged from 0.777 to 0.952. On the other hand, Table 4 illustrates that certain original scale items fell short of the required value for satisfactory convergent validity. Additional indicators of convergent validity include composite reliability [CR] and average variance extracted (AVE). Cronbach's α is determined when the items are unweighted, whereas CR is found when the items are weighted (Hair et al., 2014). This implies that composite reliability shows greater internal consistency than Cronbach's α. As stated by Hair et al. (2014), CR value of 0.7 or greater indicates excellent convergent validity and high reliability. Table 5 shows CR value for two distinct constructs. Table 5 shows the AV-adjusted scale is more reliable than the original scale.

Another technique to test convergent validity is to examine the average variance extracted [AVE] for each study

construct. AVE value should be greater than 0.5 to achieve higher convergent validity. Again, AVE value of two scales for each construct was compared. Following the vignette adjustment, AVE value for mathematical self-efficacy increased from 0.368 to 0.575, while the AVE value for math enjoyment increased from 0.615 to 0.737 (Table 5). To achieve convergent validity, CR value should be greater than the corresponding AVE value. Certainly, CR value is greater than the AVE value for each construct (see Table 5). The variation between these two values is larger for the vignette-adjusted scale. The findings of the analysis revealed that the AV approach demonstrated a favorable influence on convergent validity and reliability.

Table 5. Cronbach's α, and CR and AVE Values of Original and AV-Adjusted Scale for Two Constructs

	Original scale			*AV*-adjusted scale		
Constructs	Cronbach's α	CR	AVE	Cronbach's α	CR	AVE
Math self-efficacy	0.64	0.631	0.368	0.79	0.800	0.575
Math enjoyment	0.84	0.823	0.615	0.89	0.893	0.737

According to the findings of this study, the raw scale shows lower internal consistency; however, it can be explained by the smaller number of items on the scale. Researchers stressed that measures of internal consistency, for instance, Cronbach's α, are positively correlated with the number of items (Cortina, 1993). The greater the number of items administered, the higher the internal consistency value produced. Yet, when the raw scale is adjusted by the vignette responses, internal consistency increases from 0.64 to 0.79 for math self-efficacy scale without being affected by the smaller number of items. This result is consistent with the findings of some previous studies (Primi et al., 2016; von Davier et al., 2017; Kh. Otgonbaatar, 2021). Following this finding, this study provided another piece of evidence, which revealed that AV approach has the potential to improve internal consistency.

Furthermore, some previous studies (Chang, 1994; Preston & Colman, 2000) reported that scales with fewer response categories tend to show higher reliability values than those with more response categories. In the current study, the original scale with 5 response categories showed a reliability coefficient of 0.631 for math self-efficacy and 0.823 for math enjoyment. In contrast, *AV*-adjusted scale with 7 response categories showed a better reliability coefficient of 0.800 for math self-efficacy and 0.893 for math enjoyment, respectively.

Discriminant Validity

Discriminant validity is "… instrument's capability to differentiate or discriminate between constructs that are theoretically different" (DeVon et al., 2007, p. 5). In other words, discriminant validity measures the level of difference between the overlapping constructs. Discriminant validity can be achieved if the square root of each construct's AVE value is above 0.5 and larger than the correlation coefficient between the two constructs (Fornel and Larcker, 1971). To see the comparison, a matrix was created in which the two values can be seen simultaneously (Table 6). First, the correlation coefficient value between the original and AV-adjusted scale constructs was provided. Then, the square root of AVE values of two constructs for both scales in bold were added

on the diagonal.

Table 6. Correlation Coefficient and AVE Matrix of the Two Constructs

	Original scale		*AV*-adjusted scale	
Constructs	A	B	A	B
Math self-efficacy (A)	0.606	--	0.758	--
Math enjoyment (B)	0.645	0.784	0.214	0.858

As presented in Table 6, for the original scale, the square root of AVE value of math self-efficacy is smaller than the correlation coefficient between the two constructs. However, the value is above 0.5. For the *AV*-adjusted scale, the square root of the AVE value of both constructs is above 0.5, and larger than the two constructs' correlation coefficient. Apparently, after the vignette adjustment, the correlation coefficient between the constructs decreases while the AVE value's square root increases. The finding suggests the AV approach has an advantage in improving discriminant validity, similar to a result reported by Primi et al. (2016) in personality research. Researchers stressed that motivation and emotion in mathematics are complicated to distinguish as these constructs are conceptually overlapping (Schukajlow et al., 2017). Therefore, the finding suggested that *AV* approach can provide new insight into a conceptual overlap between emotion and motivation in mathematics education from the measurement perspective.

Conclusion

This study examined the psychometric properties of the anchoring vignette approach to measure math self-efficacy and math enjoyment. According to the findings, the following conclusions were drawn in the present study. Firstly, AV approach increased two types of reliability coefficients, including composite reliability and Cronbach's alpha, and two types of validity, such as convergent and discriminant validity. However, this study did not test criterion validity which should be examined in future studies. Secondly, the AV-adjusted scale with 7 response categories showed a better reliability coefficient than the original self-rating scale with 5 response categories. The finding encourages future studies to utilize anchoring vignettes in investigating the relationship between a number of response categories in Likert-type scales and reliability coefficient. Thirdly, AV approach has shown a positive effect on discriminant validity, which means it can be an effective measure for deciding the conceptual overlap between emotion and motivation in mathematics. In sum, the results of the present study concluded that AV approach has the potential to improve the reliability and validity of tools to measure emotion and motivation in mathematics.

In the meantime, the study investigating the effect of AV approach as a measure of emotion and motivation in mathematics contributes to existing literature, but there are still some limitations. Firstly, AV non-parametric approach utilized in this study has limitations since it considers order violation as ties. Because examining order violations as ties may lead to loss of data, parametric approaches for AV approach can be employed in future studies to deal with the disadvantage of the non-parametric approach. Secondly, the participants selected in the present study are only lower secondary grade students from public schools. As such, different findings can be

expected from a work in which wider age groups of students from different types of schools are included.

Acknowledgements

The authors thank the school principals and math teachers for allowing us to collect the data and for their cooperation. A special thanks goes to the students who participated in this study.

References

Afthanorhan, W. M. A. B. W. ., Ahmad, S. ., & Mamat, I. . (2014). Pooled Confirmatory Factor Analysis (PCFA) Using Structural Equation Modeling on Volunteerism Program: A Step by Step Approach. International Journal of Asian Social Science, 4(5), 642–653. Retrieved from https://archive.aessweb.com/index.php/5007/article/view/2663

Aybek, E. C., & Toraman, Çetin. (2022). How Many Response Categories are Enough for Likert Type Scales? : An Empirical Study Based on Item Response Theory. *International Journal of Assessment Tools in Education*, *9*(2), 534-547. Retrieved from https://www.ijate.net/index.php/ijate/article/view/81

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 307-337). Greenwich, CT: Information Age Publishing.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. Applied Psychological Measurement, 18(3), 205-215. https://doi.org/10.1177/014662169401800302

Cheeseman, J., & Mornane, A. (2014). Primary Students' Perceptions of Their Mathematics Learning. *Mathematics Education Research Group of Australasia*. https://files.eric.ed.gov/fulltext/ED572555.pdf

Chen, C., Lee, S., & Stevenson, H. (1995). Response Style and Cross-Cultural Comparisons of Rating Scales Among East Asian and North American Students. Psychological Science, 6(3),170-175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

Cortina J. M. (1993) What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, pp. 98-104.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship*, *39*(2), 155-164. https://doi.org/10.1111/j.1547-5069.2007.00161.x

Fornell, C., & Larcker, D. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. Journal of Marketing Research, 18(1), 39. https://doi.org/10.2307/3151312

Grol-Prokopczyk, H. (2014). Age and sex effects in anchoring vignette studies: Methodological and empirical contributions. In Survey Research Methods (Vol. 8, No. 1, p. 1). NIH Public Access. https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC4302337&blobtype=pdf

Grootenboer, P., & Marshman, M. (2016). Mathematics, Affect and Learning (1st ed., pp. 131-135). Springer Science + Business Media Singapore. https://doi.org/10.1007/978-981-287-679-9

Hair, J.F., Black, W. C., Babin, B. J. and Anderson, R. E. (2010). “Multivariate Data Analysis”. (7th Ed.). Prentice

Hall, Englewood Cliffs.

Hair, J. F., Hult, G.T.M., Ringle, C.M. and Sarstedt, M.M. (2014). "A Premier on Partial Least Squares Structural Equation Modeling (PLS-SEM)". SAGE Publications, California, U.S.A.

Hinz, A., Häuser, W., Glaesmer, H., & Brähler, E. (2016). The relationship between perceived own health state and health assessments of anchoring vignettes. International Journal of Clinical and Health Psychology, 16(2), 128-136. https://doi.org/10.1016/j.ijchp.2016.01.001

King, G., Murray, C., Salomon, J., & Tandon, A. (2004). Enhancing the Validity and CrossCultural Comparability of Measurement in Survey Research. American Political Science Review, 98(1), 191-207. https://doi.org/10.1017/s000305540400108x

King, G., & Wand, J. (2007). Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes. Political Analysis, 15(1), 46-66. https://doi.org/10.1093/pan/mpl011

Kitsantas, A., Ware, H. W., & Cheema, J. (2010). Predicting Mathematics Achievement from Mathematics Efficacy: Does Analytical Method Make a Difference? The International Journal of Educational and Psychological Assessment, 5, 25-44.

Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. Learning and Instruction, 15(5), 381-395. https://doi.org/10.1016/j.learninstruc.2005.07.007

Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. Labour Economics, 15(1), 96-117. https://doi.org/10.1016/j.labeco.2006.11.001

Kusmaryono, I., Suyitno, H., Dwijanto, D., & Dwidayati, N. (2019). The Effect of Mathematical Disposition on Mathematical Power Formation: Review of Dispositional Mental Functions. *International Journal of Instruction*, *12*(1), 343-356. https://doi.org/10.26711/007577152790022

Lee, J., & Stankov, L. (2013). Higher-order structure of noncognitive constructs and prediction of PISA 2003 mathematics achievement. *Learning and Individual Differences*, *26*, 119-130. https://doi.org/10.1016/j.lindif.2013.05.004

McLeod, D. (1994). Research on Affect and Mathematics Learning in the JRME: 1970 to the Present. *Journal for Research in Mathematics Education*, *25*(6), 637-647. https://doi.org/10.5951/jresematheduc.25.6.0637

Otgonbaatar, K. (2020). Examining Mathematical Creativity Among Mongolian Ninth-Grade Students Using Problem-Posing Approach. Journal of Education and Practice, 11(27), 69-75. https://doi.org/10.7176/jep/11-27-08

Otgonbaatar, K. (2021a). The development of a theoretical framework and tools to measure social and emotional skills in mathematics in the Mongolian lower secondary education (Doctoral dissertation, 広島大学). Retrieved from https://ci.nii.ac.jp/naid/500001500518/

Otgonbaatar, K. (2021b). Effectiveness of anchoring vignettes in re-evaluating self-rated social and emotional skills in mathematics. International Journal of Evaluation and Research in Education (IJERE), 10(1), 237. https://doi.org/10.11591/ijere.v10i1.20716

Paccagnella, O. (2011). Anchoring vignettes with sample selection due to non-response. *Journal of The Royal Statistical Society: Series A (Statistics In Society)*, *174*(3), 665-687. https://doi.org/10.1111/j.1467-985x.2011.00707.x

Pacheco, J. (2019). Using Anchoring Vignettes to Reevaluate the Link between Self-Rated Health Status and Political Behavior. *Journal of Health Politics, Policy and Law*, *44*(3), 533-558.

https://doi.org/10.1215/03616878-7367060

Paulhus, D. (1991). *Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes* (pp. 17-59). San Diego: Academic Press.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review, 18, 315-341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child development*, *88*(5), 1653-1670. https://doi.org/10.1111/cdev.12704

Poksinska, B., & Cronemyr, P. (2017). Measuring quality in elderly care: possibilities and limitations of the vignette method. *Total Quality Management & Business Excellence*, *28*(9-10), 1194-1207. https://doi.org/10.1080/14783363.2017.1303875

Preston, C.C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica 104, 1-15. https://doi.org/10.1016/s0001-6918(99)00050-5

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. (2016). Anchoring Vignettes. *European Journal of Psychological Assessment*, *32*(1), 39-51. https://doi.org/10.1027/1015-5759/a000336

Schukajlow, S., & Rakoczy, K. (2016). The power of emotions: Can enjoyment and boredom explain the impact of individual preconditions and teaching methods on interest and performance in mathematics? Learning and Instruction, 44, 117-127. https://doi.org/10.1016/j.learninstruc.2016.05.001

Schukajlow, S., Rakoczy, K., & Pekrun, R. (2017). Emotions and motivation in mathematics education: theoretical considerations and empirical contributions. *ZDM*, *49*(3), 307-322. https://doi.org/10.1007/s11858-017-0864-6

Skaalvik, E. M., Federici, R. A., & Klassen, R. M. (2015). Mathematics achievement and self-efficacy: Relations with motivation for mathematics. International Journal of Educational Research, 72, 129-136. http://dx.doi.org/10.1016/j.ijer.2015.06.008

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *International Journal of Academic Research in Management*, *5*.http://dx.doi.org/10.2139/ssrn.3205040

von Davier, M., Shin, H., Khorramdel, L., & Stankov, L. (2017). The Effects of Vignette Scoring on Reliability and Validity of Self-Reports. *Applied Psychological Measurement*, *42*(4), 291-306. https://doi.org/10.1177/0146621617730389

Vonkova, H., & Hrabak, J. (2015). The (in) comparability of ICT knowledge and skill self-assessments among upper secondary school students: The use of the anchoring vignette method. *Computers & Education*, *85*, 191-202. https://doi.org/10.1016/j.compedu.2015.03.003

Wadlington, E., & Wadlington, P. (2008). Helping Students With Mathematical Disabilities to Succeed. *Preventing School Failure: Alternative Education For Children And Youth*, *53*(1), 2-7. https://doi.org/10.3200/psfl.53.1.2-7

Wand, J., King, G., & Lau, O. (2016). Statistical analysis of surveys with anchoring vignettes (Version 3.0-8)

[Windows]. USA: GPL.

Weiss, S., & Roberts, R. (2018). Using Anchoring Vignettes to Adjust Self-Reported Personality: A Comparison Between Countries. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00325

West, M., Kraft, M., Finn, A., Martin, R., Duckworth, A., Gabrieli, C., & Gabrieli, J. (2016). Promise and Paradox: Measuring Students' Non-Cognitive Skills and the Impact of Schooling. *Educational Evaluation And Policy Analysis*, *38*(1), 148-170. https://doi.org/10.3102/0162373715597298

Yıldırım, S. (2012). Teacher support, motivation, learning strategy use, and achievement: A multilevel mediation model. The Journal of Experimental Education, 80(2), 150-172. http://dx.doi.org/10.1080/00220973.2011.596855

Zan, R., Brown, L., Evans, J., & Hannula, M. (2006). Affect in Mathematics Education: An Introduction. *Educational Studies in Mathematics*, *63*(2), 113-121. https://doi.org/10.1007/s10649-006-9028-2

Zimmermann, M., Bescherer, C., & Spannagel, C. (2010). A Questionnaire for Surveying Mathematics Self-Efficacy Expectations of Future Teachers.

Author Information

Khajidmaa Otgonbaatar

https://orcid.org/0000-0001-9633-626X

Mongolian National Institute for Educational Research

304, Teacher Development Building, Peace Avenue-10, Sukhbaatar district, Ulaanbaatar

Mongolia

Contact e-mail: *otgonbaatar@mier.mn*

Itgel Miyejav

https://orcid.org/0000-0001-9854-5604

Mongolian National Institute for Educational Research

304, Teacher Development Building, Peace Avenue-10, Sukhbaatar district, Ulaanbaatar

Mongolia