Detection of Differential Item Functioning for the Nine-Questions Depression Rating Scale for Thai North Dialect

Suttipong Kawilapat, Benchlak Maneeton, Narong Maneeton, Sukon Prasitwattanaseree, Thoranin Kongsuk, Suwanna Arunpongpaisal, Jintana Leejongpermpool, Supattra Sukhawaha, Patrinee Traisathit

Suttipong Kawilapat

Department of Statistics, Faculty of Science, Chiang Mai University

4 - 5 AUGUST 2021

CONTENTS

Introduction

- Depression is one of the common mental disorders and leading causes of the global disease burden such as suicide.
- In 2017, an estimated 264 million people worldwide and 2.62 million people in Thailand experienced depression.
- The prevalence of depression worldwide was estimated as 3.44% (range 2% - 6%) and as 3.09% in Thailand.

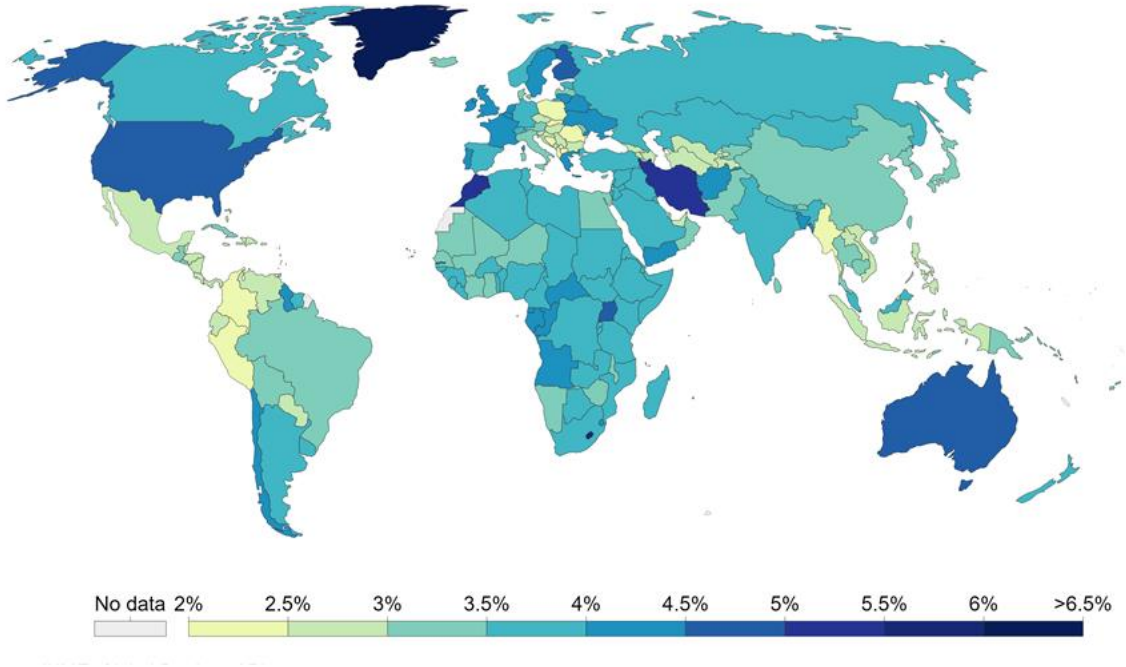

Figure 1. Prevalence of depression by country, 2017

Source: http://ourworldindata.org/mental-health

Introduction

Example of Assessment Tools

- Hamilton Rating Scale for Depression (HRSD-17)
- The Beck Depression Inventory (BDI)
- The Montgomery-Asberg Depression Rating Scale (MADRS)
- The 9-Item Patient Health Questionnaire (PHQ-9)

Introduction

Table 1. Items in Nine-Questions Depression Rating Scale (9Q)

Over the last 2 weeks, how often have you been bothered by any of the following problems? [a]	No	Yes						Score (S x F)
		Severity (S)			Frequency (F)			
		Low	Moderate	High	Several days	More than a week	Nearly every day	
1. Depressed mood	0	1	2	3	1	2	3	
2. Markedly diminished interest or pleasure	0	1	2	3	1	2	3	
3. Insomnia or hypersomnia	0	1	2	3	1	2	3	
4. Fatigue or loss of energy	0	1	2	3	1	2	3	
5. Significant weight loss or gain	0	1	2	3	1	2	3	
6. Feeling of worthlessness or excessive or inappropriate guilt	0	1	2	3	1	2	3	
7. Diminished ability to think or concentrate, or indecisiveness	0	1	2	3	1	2	3	
8. Psychomotor agitation or retardation	0	1	2	3	1	2	3	
9. Recurrent thoughts of death, recurrent suicidal ideation, or a suicide attempt	0	1	2	3	1	2	3	

[a] Depressive symptoms according to the fifth edition of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-5)

Introduction

- The observed scores during measurement can be different between groups due to differences in the true trait ability or differences among the characteristics of the respondents.
- Different probabilities of response to an item among respondents with the same underlying trait score is defined as differential item functioning (DIF), the presence of which may compromise comparisons across subgroups and can lead to misleading results (Crane et al., 2004).
- The traditional calculation of the measurement score when not considering DIF might not be proper when the DIF occurs.

Objectives

The aim of this study is to determine the presence of DIF in responses to depression assessment tools and related factors pertaining to the Thai population.

Literature Review

- DIF occurs when participants from different demographic groups (e.g., gender, age) with the same underlying trait score have a different probability of responding to an item.
- There are 2 types of DIF (Crane et al., 2006):
  - **Non-uniform DIF (NUDIF)**
    A statistically significant interaction between the trait level and the demographic variable (effect modification)
  - **Uniform DIF (UDIF)**
    The difference between the strength of the relationship between the ability and item responses in a model with and without the demographic variable for each item (confounding)

Literature Review

The ordinal logistic regression (OLR) technique is an approach based on traditional sum score to determine DIF for polytomous items. The following ordinal response models are fitted prior to exploring both UDIF and NUDIF:

$$f(\text{item response}) = \beta_0 + \beta_1\theta \quad (1)$$

$$f(\text{item response}) = \beta_0 + \beta_1\theta + \beta_2 X \quad (2)$$

$$f(\text{item response}) = \beta_0 + \beta_1\theta + \beta_2 X + \beta_3(\theta X) \quad (3)$$

Likelihood ratio test or relative difference of β_1 (≥10%) between (1) and (2)

Likelihood ratio test between (2) and (3)

where β_0 = the intercept coefficients

β_1 = the regression coefficients of the trait level θ

β_2 = the regression coefficients of variable X

β_3 = the regression coefficients of the interaction between trait level θ and variable X

Literature Review

- An approach based on the item response theory (IRT) is proposed as an alternative method to determine DIF accounting for underlying trait in addition of sum score.
- The baseline IRT models are fitted for all items and then compared to the other model with varied discrimination and threshold parameters between the reference and focal groups for each item.
- A comparison of models is performed using the likelihood ratio test between the baseline and constrained model indicating the presence of DIF between the groups (Raykov and Marcoulides, 2018).

Literature Review

Table 2 . Common item response theory models for polytomous items

Model	Parameters	Characteristics
Graded Response	• Threshold • Discrimination	• Ordered responses • Discrimination varies across items
Rating Scale	• Threshold	• Equal step threshold parameters for all items
Partial Credit	• Threshold	• Different step threshold parameters for all items
Generalized Partial Credit	• Threshold • Discrimination	• Variation of Partial Credit Model with discrimination varying across items.

Source: Cappelleri (2014)

Literature Review

The **generalized partial credit model (GPCM)**, which is an IRT model for polytomous items, was preferable for estimating the IRT parameters in this study (Muraki, 1992; Edelen and Reeve, 2007; Nering and Ostini, 2010)

$$P_{ik}(\theta) = \frac{\exp\left[\sum_{k=1}^{m} a_i\left(\theta - b_{ik}\right)\right]}{\sum_{h}^{m-1} \exp\left[\sum_{k=1}^{h} a_i\left(\theta - b_{ik}\right)\right]}$$

where $P_{ik}(\theta)$ = the probability of responding to item i in category k (k = 0, 1, … , m)

a_i = the discrimination parameter of item i

b_{ik} = the threshold parameter for item i in category k

Literature Review

- Broekman et al. (2008) conducted a study using Multiple Indicator, Multiple Cause model (MIMIC) to examine DIF across characteristics for items in the Geriatric Depression Scale-15 (GDS-15) among the elderly in Singapore. They found significant DIF associated with age, gender, ethnicity, and chronic illness for 8 items.

- Cameron (2013) conduct a study using Mantel's χ^2, Liu-Agresti cumulative common odds ratio, and standardized Liu-Agresti cumulative common odds ratio to examine DIF across age, gender, and educational background in depression assessment tools i.e., the Patient Health Questionnaire (PHQ-9) and the Hospital Anxiety and Depression Scale (HADS). They found that age-related DIF on 3 items in PHQ-9 and 2 items in HADS Depression subscale.

Literature Review

- Patel et al. (2014) conducted a study among patients undergone treatment for painful conditions in the emergency department in the United States and found that female patients presented with higher scores for stress and anxiety than male.
- Jiraniramai et al. (2021) using Rash analysis approach to examine the DIF of PHQ-9 items among health care workers in Chiang Mai, Thailand and found no significant DIF across age, gender, educational level, and alcohol consumption in any item.

Methodology

Setting and Participants

- We used secondary data from a study on the criterion-related validity of a revised 9Q in the northern Thai dialect comprising 1,527 individuals from the northern region of Thailand.
- This questionnaire was translated from the central Thai dialect (Kongsuk et al, 2018).
- Participants who did not complete all items in the assessment were excluded.

Ethical Approval

- This study using de-identified data from the primary study approved by the Ethical Committee, Phra Si Maha Phot Psychiatric Hospital, Ubon Ratchathani, Thailand.

Methodology

Statistical Analyses

- The demographics of the participants are reported as frequencies and percentages for categorical variables and as medians and interquartile ranges (IQRs) for the continuous variables.
- The OLR approach was performed using the Stata "**DIFDETECT**" command (Crane et al., 2006).
- The generalized partial credit model (GPCM) was preferable for estimating the IRT parameters in this study.
- *p*-value <0.05 was considered to be significant in all analyses.
- All analyses were performed using Stata 17 (StataCorp, College Station, Texas, USA).

Results and Discussion

Table 3. Demographic characteristics of the participants (N=1,475).

Characteristic	n (%) or Median [IQR]
Gender	
Male	493 (33.4%)
Female	982 (66.6%)
Age	45 [33–57]
13–18	120 (8.1%)
18–59	1,086 (73.6%)
60+	269 (18.3%)
Underlying disease (n = 1,459)	
No	931 (63.8%)
Yes	528 (36.2%)
Income (baht/month) (n = 1,453)	
<5,000	759 (52.2%)
5,000–10,000	442 (30.4%)
>10,000	252 (17.3%)

- The majority were female (66.6%), adult (73.6%), did not have any underlying diseases (63.8%), and had an income per month of less than 5,000 baht (52.2%).
- The median age of the participants was 45 (IQR 33–57) years old.
- Some participants did not complete their information about underlying disease and income on patient report form

Results and Discussion

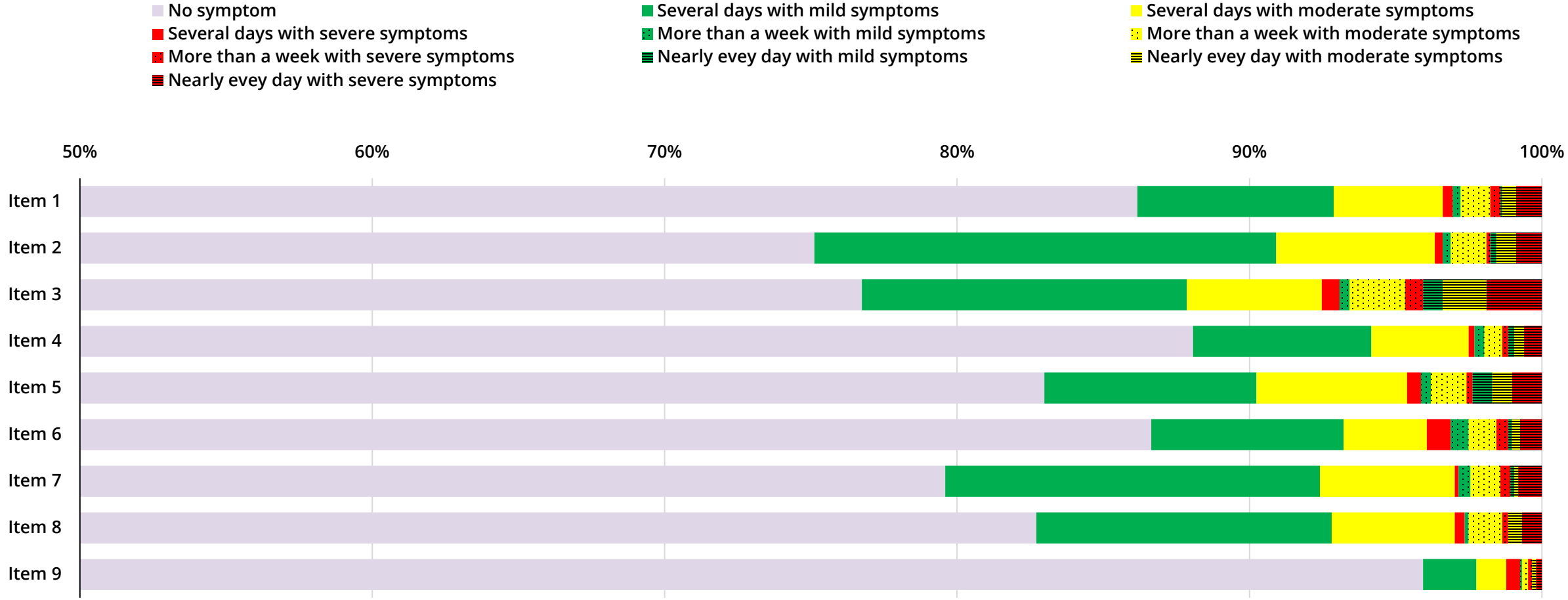

Figure 2. Item endorsement by the participants.

Results and Discussion

Table 4. Estimated parameter based on item response theory using graded response model.

Item	IRT Parameters from the GPCM									
	a_i	$b_{i(11)}$	$b_{i(12)}$	$b_{i(13)}$	$b_{i(21)}$	$b_{i(22)}$	$b_{i(23)}$	$b_{i(31)}$	$b_{i(32)}$	$b_{i(33)}$
1. Depressed mood	2.533	1.508	1.442	2.399	1.659	1.126	2.140	2.393	1.046	1.642
2. Markedly diminished interest or pleasure	1.885	1.069	1.549	2.910	1.489	0.793	2.829	1.509	1.145	1.713
3. Insomnia or hypersomnia	0.376	5.117	2.623	5.924	2.334	-3.624	4.126	0.952	-0.762	0.778
4. Fatigue or loss of energy	1.837	1.890	1.484	2.957	1.295	1.338	2.322	1.781	1.563	1.591
5. Significant weight loss or gain	0.428	5.734	1.188	6.247	1.701	-1.889	5.430	-1.448	1.456	0.587
6. Feeling of worthlessness or excessive or inappropriate guilt	1.167	2.473	1.617	2.318	1.681	1.174	2.363	2.649	1.178	0.971
7. Diminished ability to think or concentrate, or indecisiveness	1.276	1.650	1.695	4.019	0.590	0.851	2.514	2.441	1.787	0.450
8. Psychomotor agitation or retardation	1.287	1.889	1.608	3.236	2.181	-0.076	3.021	1.091	NA	1.547
9. Recurrent thoughts of death, recurrent suicidal ideation, or a suicide attempt	2.387	2.372	1.679	1.962	2.564	1.367	2.070	1.979	NA	1.905

Abbreviations: IRT, the item response theory; GPCM, the generalized partial credit model; a_i, the discrimination parameter of item i; $b_{i(fs)}$, the threshold parameter of item i with frequency f and severity s.

Results and Discussion

- More than 80% of the participants had no symptoms related to depression within the previous two weeks, except for items 2, 3, and 7.
- Item 3 had the highest endorsement rate of having severe symptoms nearly every day.
- Almost all of the participants (96%) did not report thoughts of physical self-harm or suicide (item 9).
- The highest discrimination parameter value was obtained for item 1, followed by items 9, 2, and 4, all of which are the most related to depression.
- The discrimination parameter values for items 3 and 5 were lower than the others.

Results and Discussion

Table 5. Detecting the presence of DIF using the OLR approach.

Item	Gender			Age			Underlying Disease			Income		
	Non-uniform	Uniform		Non-uniform	Uniform		Non-uniform	Uniform		Non-uniform	Uniform	
	(*p*)	(%)	(*p*)	(*p*)	(%)	(*p*)	(*p*)	(%)	(*p*)	(*p*)	(%)	(*p*)
1	0.325	0.0001	0.073	0.995	0.0024	0.386	**0.009**	-0.0015	0.838	0.189	-0.0031	0.674
2	**0.016**	0.0060	**0.002**	0.071	0.0264	**<0.001**	**<0.001**	0.0262	**0.006**	0.861	0.0024	0.303
3	0.665	-0.0006	0.072	0.228	0.0014	**0.009**	0.217	-0.0231	**0.003**	0.067	-0.0040	**0.007**
4	0.071	-0.0003	0.895	**0.004**	0.0113	**0.002**	0.920	-0.0118	**0.024**	0.055	0.0004	**0.021**
5	0.257	0.0005	0.734	**0.023**	0.0010	0.919	**0.009**	-0.0169	0.121	0.574	-0.0014	0.148
6	0.829	0.0016	0.310	0.363	0.0254	**<0.001**	0.459	0.0146	0.215	0.194	-0.0026	0.359
7	0.299	0.0003	0.895	**0.010**	0.0110	**0.005**	0.165	0.0028	0.716	0.230	-0.0035	0.529
8	0.730	0.0065	0.842	0.422	-0.0008	0.779	**0.004**	-0.0199	**0.001**	0.416	-0.0042	0.684
9	0.141	-0.0005	0.837	0.564	-0.0047	0.847	0.431	-0.0240	0.068	0.410	0.0124	0.024

The items in bold indicate the presence of DIF (differential item functioning).

Results and Discussion

Table 6. Detecting the presence of DIF using the IRT approach.

Item	Gender		Age		Underlying disease		Income	
	Non-uniform (*p*)	Uniform (*p*)	Non-uniform (*p*)	Uniform (*p*)	Non-uniform (*p*)	Uniform (*p*)	Non-uniform (*p*)	Uniform (*p*)
1	0.408	0.320	**0.018**	**0.016**	0.843	0.836	0.546	0.420
2	0.094	0.066	**0.001**	**0.004**	**0.005**	**0.008**	0.241	0.229
3	0.112	0.242	0.126	0.126	**0.049**	**0.031**	0.100	0.116
4	0.687	0.586	**0.004**	**0.002**	0.599	0.520	0.775	0.734
5	**0.003**	**0.017**	0.219	0.366	0.051	**0.034**	0.179	0.150
6	0.314	0.240	0.351	0.261	0.240	0.196	0.194	0.185
7	0.868	0.802	**0.035**	0.185	**0.041**	0.255	0.653	0.653
8	**0.006**	**0.003**	0.161	0.086	**0.003**	**0.003**	0.084	0.125
9	0.568	0.761	0.838	0.753	0.568	0.452	0.709	0.713

The items in bold indicate the presence of DIF (differential item functioning).

Results and Discussion

OLR approach

- DIF between the characteristics was present in all of the studied variables:
  - gender → item 2
  - age → items 2, 3, 4, 5, 6, 7
  - underlying disease → items 1, 2, 3, 4, 5, 8
  - income → items 3 and 4
- NUDIF was present across the groups for all of the characteristics except income.
- UDIF was present across all of the characteristics.

IRT approach

- DIF between the characteristics was present in all of the studied variables except income:
  - gender → item 5, 8
  - age → items 1, 2, 4, 7
  - underlying disease → items 2, 3, 5, 7, 8
- NUDIF was present across the groups for all of the characteristics except income.
- UDIF was present across all of the characteristics except income.

Results and Discussion

- Findings from previous studies suggest that the IRT technique may reveal additional information about the actual level of the underlying trait compared to the observed score (Reise and Haviland, 2005; Snitz et al., 2012; Gorter et al., 2015; Saracino et al., 2020).
- Since DIF occurs from the effect modification between the characteristics and the trait level or the difference in the strength of the relationship between the ability and item responses, the inconsistency in the OLR approach might have resulted from not accounting for the different discrimination and threshold parameters related to depression for each item.
- Therefore, the IRT approach is probably more appropriate for examining the DIF in polytomous items.

Results and Discussion

Item	Gender	
	Non-uniform (*p*)	Uniform (*p*)
1	0.408	0.320
2	0.094	0.066
3	0.112	0.242
4	0.687	0.586
5	**0.003**	**0.017**
6	0.314	0.240
7	0.868	0.802
8	**0.006**	**0.003**
9	0.568	0.761

- Both NUDIF and UDIF were present for gender in 2 items
  - item 5: significant weight loss or gain
  - item 8: psychomotor agitation or retardation
- These significant DIF values between gender might have resulted from the natural difference concerning gender on perception or concern about psychological issues.
- The study of Patel et al. (2014) indicated that female patients presented higher scores for stress and anxiety than male.
- The study of Udo et al. (2014) also showed that stressful life events are associated with an increase in BMI in females.

Results and Discussion

Item	Age	
	Non-uniform (*p*)	Uniform (*p*)
1	**0.018**	**0.016**
2	**0.001**	**0.004**
3	0.126	0.126
4	**0.004**	**0.002**
5	0.219	0.366
6	0.351	0.261
7	**0.035**	0.185
8	0.161	0.086
9	0.838	0.753

- DIF were present for age groups in 4 items
  - item 1: depressed mood
  - item 2: markedly diminished interest or pleasure
  - item 4: fatigue or loss of energy
  - item 7: diminished ability to think or concentrate, or indecisiveness
- The 9Q items with age-related DIF found in this study was consistent with the study of Cameron (2013).
- Applying an appropriate tool to measure depression according to differences of age (the Children's Depression Inventory or the Geriatric Depression Scale-15, might have resulted in a reduction in bias in the assessments.

Results and Discussion

- In addition to gender and age, the items related to a feeling of worthlessness and loss of energy among the elderly presented DIF in the findings from a previous study conducted among the elderly using the GDS-15 across chronic illness groups (Broekman et al., 2008).
- The DIF across illness was also presented in our study for the item related to loss of energy.
- This might have been because of the impact of different illnesses leading to a difference in fatigue level across participants with and without underlying illnesses.
- A recent study of Jiraniramai et al. (2021) did not find the DIF in any items across age, gender, education and alcohol consumption. They suggested that it might be related with the no to low level of depression for the health care workers.

Results and Discussion

Limitations & Suggestions

- The lower number of participants across other interesting variables such as nationality, ethnicity, educational background, or occupation. A further study with a larger sample size should be conducted to determine DIF in other variables and confirm the findings presented in the present study.
- A recent study on the impact of somatic symptoms on PHQ-9 scores found that several items showed DIF with respect to disease-specific severity, however, the salient DIF was present in very few patients (Katzan et al., 2021). Considering for the impact of DIF related to characteristics could be useful in further study.
- In addition, other approaches toward determining the DIF for polytomous items should be considered.

Conclusion

- In this cross-sectional study to determine the presence of DIF in the responses to the 9Q tool for depression severity assessment among the northern Thai population, DIF was found in the responses for several items according to the participants' characteristics including gender, age, and underlying disease except item 6 and 9.
- The findings from our study suggest that the IRT approach should be used to determine DIF for polytomous items.
- In addition, accounting for the difference between the characteristics of participants might reduce the bias in the scoring or assessment of depression severity.

Acknowledgment

We would like to thank the physicians, nurses, medical staffs, and all participants who involved in this study. A primary study on validity of 9Q among northern Thai population was funded by a grant from Mental Department of Mental Health, Ministry of Public Health. This study was partially supported from Chiang Mai University, and Department of Statistics, Faculty of Science, Chiang Mai University.

References

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5®): American Psychiatric Publishing.

Broekman, B.F., Nyunt, S.Z., Niti, M., Jin, A.Z., Ko, S.M., Kumar, R., et al. (2008). Differential item functioning of the Geriatric Depression Scale in an Asian population. *J Affect Disord*, 108(3):285–290.

Cameron, I. M., Crawford, J. R., Lawton, K., and Reid, I. C. (2013). Differential item functioning of the HADS and PHQ-9: An investigation of age, gender and educational background in a clinical UK primary care sample. *Journal of Affective Disorder*, 147:262–268.

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical therapeutics*, *36*(5), 648–662.

Crane, P. K., van Belle, G., and Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Stat Med*, 23:241–256.

Crane, P. K., Gibbons, L. E., Jolley, L., and van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Med Care*, 44(11 Suppl 3):S115–S123.

Edelen, M. O., and Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res*, 16(Suppl 1):5–18.

Gorter, R., Fox, J. P., and Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Med Res Methodol*, 15:55.

Jiraniramai, S., Wongpakaran, T., Angkurawaranon, C., Jiraporncharoen, W., and Wongpakaran, N. (2021). Construct Validity and Differential Item Functioning of the PHQ-9 Among Health Care Workers: Rasch Analysis Approach. *Neuropsychiatric disease and treatment*, 17:1035–1045.

Johnson, J. G., Harris, E. S., Spitzer, R. L., and Williams, J. B. (2002). The patient health questionnaire for adolescents. *Journal of Adolescent Health*, 30, 196–204.

Katzan, I.L., Lapin, B., Griffith, S., Jehi, L., Fernandez, H., Pioro, E., et al. (2021). Somatic symptoms have negligible impact on Patient Health Questionnaire-9 depression scale scores in neurological patients. *Eur J Neurol*, 28:1812–1819.

Kongsuk, T., Arunpongpaisal, S., Janthong, S., Prukkanone, B., Sukhawaha, S., and Leejongpermpoon, J. (2018). Criterion-Related Validity of the 9 Questions Depression Rating Scale revised for Thai Central Dialect. *Journal of the Psychiatric Association of Thailand*, 63:321–334.

References

Muraki E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, 16:159–176.

Nering, M.L., and Ostini, R., eds. (2010). Handbook of Polytomous Item Response Theory Models. New York: Routledge Academic.

Patel, R., Biros, M.H., Moore, J., and Miner, J.R. (2014). Gender differences in patient-described pain, stress, and anxiety among patients undergoing treatment for painful conditions in the emergency department. *Acad Emerg Med*, 21:1478–1484.

Raykov, T. , and Marcoulides, G. A. (2018). A course in item response theory and modeling with Stata. College Station, TX: Stata Press College Station.

Reise, S. P., and Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *J Pers Assess*, 84:228–238.

Ritchie, H., and Roser, M., (2018). Mental Health: Our World in Data. Available from: https://ourworldindata.org/mental-health. Cited 10 Mar 2019.

Saracino, R. M., Aytürk, E., Cham, H., Rosenfeld, B., Feuerstahler, L. M., and Nelson, C. J. (2020). Are we accurately evaluating depression in patients with cancer? *Psychol Assess*, 32:98–107.

Sheikh, J. I., and Yesavage, J. A. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontologist*, 5:165–173.

Snitz, B. E., Yu, L., and Crane, P. K., (2012). Chang CC, Hughes TF, Ganguli M. Subjective cognitive complaints of older adults at the population level: an item response theory analysis. *Alzheimer Dis Assoc Disord*, 26:344–351.

Trangkasombat, U., and Likanapichitkul, D. (1997). The Children's Depression Inventory as a screen for depression in Thai children. *J Med Assoc Thai*, 80:491–499.

Udo, T., Grilo, C. M., and McKee, S. A. (2014). Gender differences in the impact of stressful life events on changes in body mass index. *Preventive Medicine*, 69:49–53.

THANK YOU