ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 03, pp.27120-27131, March, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

EXAMINING PSYCHOMETRIC PROPERTIES OF THE KUTCHER ADOLESCENT DEPRESSION SCALE (KADS-11) USING MULTIDIMENSIONAL ITEM RESPONSE THEORY

***,[1]Mahnaz Shojaee, [1]Okan Bulut and [2]Mehrdad Shahidi**

[1]Centre for Research in Applied Measurement and Evaluation, University of Alberta
11210 87 Ave NW Edmonton, AB T6G 2G5, Canada
[2]Mount Saint Vincent University, 166 Bedford Highway, Halifax, Nova Scotia, B3M 2J6, Canada

ARTICLE INFO

Article History:
Received 07th December, 2015
Received in revised form
20th January, 2016
Accepted 27th February, 2016
Published online 16th March, 2016

Key words:

Item response theory, Depression,
Graded Response Model,
Multidimensionality,
Kutcher Adolescent Depression Scale.

ABSTRACT

Kutcher Adolescent Depression Scale-11 Items (KADS-11) is a diagnostic instrument measuring depression and suicidal thoughts in adolescents and young adults. Some characteristics of KADS-11 such as ease of administration, treatment sensitivity, and the ability to distinguish comorbid symptoms motivated the researchers to examine the psychometric properties of the scale by utilizing the multidimensional form of Graded Response Modeling in order to find the relationship between item responses and the latent trait. Results indicated that most of the items provided the maximum amount of information about participants' depression, and also two extracted factors (Core Depressive factor and Suicidal and Physical factor) can explain 55.20% of total variance of KADS-11.

Citation: **Mahnaz Shojaee, Okan Bulut and Mehrdad Shahidi, 2016.** "Examining psychometric properties of the Kutcher adolescent depression scale (KADS-11) using multidimensional item response theory", *International Journal of Current Research*, 8, (03), 27120-27131.

INTRODUCTION

Depression is one of the prevalent mental disorders across Canada (Centre for Addiction and Mental Health-CAMH, 2011; Canadian Mental Health Association-CMHA, 2013) and around the world (World Health Organization, 2015). This disorder, which is characterized by several symptoms such as depressed mood or irritable, loss of interest, low morale, a sense of hopelessness, lack of happiness and self-worth, withdrawal, psychomotor retardation and other symptoms (Steptoe, Tsuda, Tanaka, and Wardle 2007; Walkiewicz, Tartas, Majkowicz and Budzinski, 2012), was predicted to be the second most common mental health disorder in Canada by 2020 (WHO, 2002 cited in CAMH, 2011). Since it is estimated that 70% of mental health problems have their onset during childhood and adolescence (CAMH, 2015) and young adults are at risk for developing depression and other types of mental health difficulties (CMHA, 2013; Ialomiteanu, Hamilton, Adlaf, and Mann, 2014), most clinicians, practitioners and counsellors endeavor to screen and diagnose depression and other types of mental health problems by relying on reliable and valid tools for further therapeutic intervention. Amongst various depressive tools, Hamilton Rating Scale for Depression (HAM-D or HRSD- Bech, Paykel, Sireling and Yiend, 2015; Broen *et al*., 2015; Hamilton, 1960), Beck Depression Inventory (BDI-Beck, Ward, Mendelson, Mock, and Erbaugh, 1961; Dere, Watters, Chee-Min Yu, Bagby, Ryderand Harkness, 2015; Lahlou-Laforêt, Ledru, Niarra, and Consoli, 2015), Inventory of Depressive Symptomatology (IDS- Rush, Giles, Schlesser, Fulton, Weissenburger, and Burns, 1986), Quick Inventory of Depressive Symptomatology (QIDS-Reilly, MacGillivray, Reid, and Cameron, 2015; Rush, 2003), Montgomery–Asberg Depression Rating Scale (MADRS-Carneiro, Fernandes, and Moreno, 2015; Montgomery, and Asberg, 1979), Zung Self-Report Depression Scale (Zung SDS – Shahidi and Shojaee, 2014; Zung, 1965), Dimensions of Depression Profile for Children and Adolescents (DDPCA-Qualter, Brown, Munn, and Rotenberg, 2010), 21–item Children's Depression Inventory (CDI- Brooks and Kutcher, 2001), Center for Epidemiology Depression Scale (CES-D-Szumilas, Kutcher, LeBlanc, and Langille, 2010), Hospital Anxiety and Depression Scale (HADS-FC- Roberge, Dore, Menear, Chartrand, Ciampi, Duhoux, and Fournier, 2013) and

**Corresponding author: Mahnaz Shojaee,*
Center for Research in Applied Measurement and Evaluation, University of Alberta 11210 87 Ave NWEdmonton, AB T6G 2G5, Canada.

Kutcher Adolescent Depression Scale-11 Items (KADS-11) (Brooks and Kutcher, 2001; Shahidi and Shojaee, 2014) are seemingly the popular scales that are used in most clinical settings. A review of studies about these scales and other depressive tools revealed that such tools should possess at least the following characteristics for assessing disorders: 1) ease of administration;2) adequate number of items; 3) the ability to distinguish comorbid symptoms; 4) specificity and sensitivity; 5) purpose: screening or diagnosis; 6) ability to measure change over time (treatment sensitivity); 7) internal reliability and validity; 8) developmentally appropriate; and 9) self-report rather than clinician administered (Brooks *et al*., 2003; Levine, 2013; Roberge, *et al*., 2013; Trujols, Feliu-Soler, Diego-Adeliño, Portella, Cebrià, Soler, Puigdemont, Álvarez, and Perez, 2013).

In addition to these characteristics and based on the nature of diagnostic tools (e.g., unidimensional or multidimensional), it is important whether such tools fulfil certain psychometric requirements or not (Ackerman, Gierl and Walker, 2003; Sheng and Wikle, 2007). The psychometric requirements for diagnostic and screening scales have been argued in two major theories, Classical Test Theory (CTT) and Item Response Theory (IRT). These theories focus on the measurement in social and also natural sciences. Measurement, which is the process of attributing numbers to the variables (Kerlinger, 1986) to quantify and assess them, is a central concept in the history of scientific progress and discoveries (Popper, 1994, 2005). This process is extremely intricate when psychologists attempt to measure traits that are naturally *subjective* rather than *objective* (Cronbach, and Meehl, 1955; Hooman, 2007; Hooman, 2010; Kane, 1992). The major questions in this process revolve around the reasonable degree of reliability, validity, consistency, stability, accuracy, predictability, and the congruency of tools or instruments by which the process of attributing numbers to subjective variables is accomplished. To provide a rational answer to these problems, the developers of psychological scales endeavor to minimize measurement errors and meet the psychometric requirements. Accordingly, they attempt to construct the tools for their own desired qualities by choosing the most appropriate elements such as items and choices (dichotomous or polytomous choices) for the tools (Güler, Uyanık, and Teker, 2014; Zoghi, and Valipour, 2014). Regardless of these endeavors, the question is to what extent such psychological diagnostic or screening scales may meet the psychometric requirements.

Reviewing the history of above noted scales revealed that most of them just meet some psychometric requirements or cover a few of aforementioned characteristics. For example, since Hamilton (1960) introduced HAM-D to practitioners, HAM-D remained unchanged for many years (Broen *et al*., 2015). HAM-D needed to have replication many times particularly because the scale was initially developed for assessing the changes in depressive symptoms over time in psychiatric population (Broen *et al*., 2015). Such problems may also be found for other scales such as QIDS (Reilly *et al*., 2015) or Zung SDS (Shahidi and Shojaee, 2014); although, they have been highly used in clinical settings (Ghanei, Golkar, and Edalat Aminpour, 2014; Makaremi, 1992; Reilly, *et al*., 2015; Shojaeizadeh and Rasafiyani, 2001). Additionally, most psychological scales have been developed based on CTT such as HAM-D, Geriatric Depression Scale, Depression Anxiety Stress Scales-21 and other scales (Allqaier, Pietsch, Fruhe, Sigl-Glockner, and Schulte-Kome, 2012; Bottesi, Ghisi, Altoè, Conforti, Melli, and Sica, 2015; Carneiro, Fernandes, and Moreno, 2015; Guerra, Ferri, Llibre, Prina, and Prince, 2015). However, some other scales have been developed using both CTT and IRT to determine the major psychometric properties of and requirements for different clinical and achievement scales (Barthel, Barkmann, Ehrhardt, and Bindt, 2014; Güler, *et al*., 2014; Zoghi, and Valipour, 2014; Rubio, Aguado, Hontangas, and Hernindez, 2007). Based on the aforementioned characteristics of diagnostic tools, it was argued that amongst such depression tools, KADS-11 may cover most of psychometric requirements (Brooks and Kutcher, 2001; Brooks, *et al.,* 2003; LeBlanc, Almudevar, Brooks, and Kutcher 2002). This scale, which was initially developed for the monitoring of treatment and assessing the changes of symptoms over time (Brooks and Kutcher, 2001), has recently caught the clinicians' attention whether it is an appropriate depression tool for clinicians or not (LeBlanc, *et al*., 2002; Shahidi and Shojaee, 2014). Although the first studies of this scale claimed that KADS-11 is a unidimensional self-report tool addressing some of above noted characteristics (Brooks and Kutcher, 2001; Brooks, *et al*., 2003; LeBlanc, *et al*., 2002), the most recent study showed that this scale is not unidimensional (Shahidi and Shojaee, 2014).

The first studies on KADS-11 also revealed that the scale can be used not only to identify and diagnose depressed adolescents; the scale is also sensitive to treatment effects. Similar to many other depressive scales, KADS-11 is easy and time efficient to be administrated. Previous analyses based on CTT indicated that KADS-11 is a reliable and valid tool for measuring depression (Brooks and Kutcher, 2001; Brooks *et al*., 2003; Shahidi and Shojaee, 2014). In addition to the above noted characteristics, the KADS-11 is developmentally appropriate that was derived from the core symptoms of adolescent depression, and it also measures the severity of those symptoms (Brooks *et al*., 2003). Thus, it may be a useful clinical tool in the diagnosis and management of adolescent depression. Comparing with Dimensions of Depression Profile for Children and Adolescents (DDPCA- Qualter, *et al*., 2010), 21-item BDI, CES-D, HADS-FC and some other depression or mental health scales, the KADS-11 has some remarkable advantages. For example, DDPCA (Qualter, *et al*., 2010) is a successful tool at identifying individuals who are at high risk of suicide (Qualter, *et al*., 2010), but it has low sensitivity to the change of symptoms during the treatment. In relation to 21-item BDI, which is made up of numerous items and takes a relatively long time to complete, KADS-11 is an easy and time efficient tool to administrate. Other scales (such as CDI, CES-D; and HADS-FC) have either low discriminative validity in adolescent depression (Brooks and Kutcher, 2001) or low sensitivity to age and developmental trajectories (Roberge, *et al*., 2013). Although these characteristics of KADS-11make the scale distinguished, its psychometric properties to show the minimum error in assessing depression should be re-examined using both CTT and IRT. Using both CTT and IRT in an integrative way to explore, develop and to maximize the applicability of scales is recently central to psychometric

studies (Barthel, *et al*., 2014; Güler, *et al*., 2014; Rubio, *et al*., 2007; Zoghi, and Valipour, 2014). These theories can help clinicians and educators to introduce the most appropriate tools to their societies and targeted population. Although the most fundamental assumptions of CTT − such as true scores of the examinees are different from IRT's assumptions in which the unidimensionality is central – both theories have comparable facts based on some recent reliable studies (Güler, *et al*., 2014; Rubio, *et al*., 2007; Sebille, *et al*., 2010; Zoghi and Valipour, 2014).

This type of validating scales not only provides an opportunity to compare the potentials of IRT and CCT, but it also provides the clinicians, practitioners and educators with the most reliable psychometric properties of a given tool. Such characteristics can be found in the recent studies of clinical instruments (Barthel, *et al*., 2014; Mead, and Meade, 2010; Rubio, *et al*., 2007; Sebille, *et al*., 2010). However, since the majority of psychological scales are multidimensional, the sensitivity of using IRT is in the first priority of validating a given multidimensional scale (Bulut, 2014). No matter which of these theories is used to develop a clinical test, researchers usually pursue a set of stages which start from the preparation of test specifications and end with the distribution of the test and its manuals (see Table 1). In regard to these stages, and to explore the strengths and weaknesses of KADS-11, the current study was focused on the psychometric properties of KADS-11. Accordingly, the following questions were addressed: 1) what are the item parameters of KADS-11 based on CTT and IRT? 2) Are the estimated parameters in IRT comparable with estimated parameters in CTT for KADS-11? 3) Does the GRM in IRT have a better fit to the data compared to CTT for KADS-11? 4) Is there any advantage for the scale to be analyzed based on multidimensional item response theory (MIRT)?

Method

Sample

This study involved 300 students who were randomly selected by using the systematic multistage random sampling method from Islamic Azad University-Tehran Central in Iran. Of this group of participants, 277 students completed the scale and 23 students did not finish all of the items on the scale. 22 (8%) individuals were males and 255 (92%) were females. The gender ratio of samples aligned with the nature of university (Valiaser Campus) population in which female students consist of approximately 80% of the total student population. The mean age of sample students was 22.8 years (SD = 4.38). All students were studying either psychology or social sciences.

Instrument

The KADS-11 is an eleven-item, self-report instrument, which was initially studied in a Canadian population (Brooks and Kutcher, 2001; Brooks, *et al*., 2003; LeBlanc, *et al*., 2002). This scale was introduced for clinical practice as a sensitive and specific instrument to aid in diagnosis and monitoring the change in severity of symptoms during the course of treatment. The KADS-11 was created using youth friendly language. It is easily and quickly completed, diagnostically valid and demonstrates reasonable reliability (Brooks and Kutcher, 2001; Brooks *et al*., 2003). The items in the KADS-11 were constructed on the basis of core symptoms of depression that measures the frequency of depressive symptoms (Brooks *et al*., 2003). The instrument consists of items with an ordinal and polytomous scoring scale, which includes the category 0 (hardly ever), 1 (much of the time), 2 (most of the time), to the category 3 (all of the time).

Since usually cutscore has been used to determine the sensitivity and severity of depression classifications (Brooks *et al*., 2001; LeBlanc *et al*., 2002), in this study the Z score ($\mu = 0$, SD = 1) was preferred to determine the degree to which participants show different levels of depression. This method provided the researchers with cut-score for estimating normality and the severity of depression. The procedure revealed that the participants who received Z score below 1 ($Z < 1$) were not depressed, who received Z score between 1 and 2 were sensitive to depression, and who received upper than 2Z score were depressed.

Procedure

Since the scale was originally in English language, the instrument was translated and back translated into Persian by two of the researchers and was evaluated by a blinded language specialist. After being confirmed about the translation, all participants completed KADS-11 in group sessions in their usual classes based on the sampling method described above. Administration of the instrument was counterbalanced. Participants took approximately 5 to 15 minutes to fill in the scale. After collecting the data, R and SPSS 22 were used to analyze the data.

Data Analysis

A review of psychological scales revealed that psychological instruments measure either one target trait, such as Procrastination Scale (Lay, 1986), Current Thoughts Scale (Heatherton and Polivy, 1991) and Zung Self-Rating Depression Scale (SDS- Carroll, Fielding, and Blashki, 1973; Zung, 1965) or they measure more than one dimension such as Self-Regulation Questionnaire (SRQ- Neal and Carey, 2004; Neal and Carey, 2005) or NEO Personality Inventory (Costa and McCrae, 1985; Costa, Terracciano, and McCrae, 2001). Accordingly, since the KADS-11 was demonstrated to be a multidimensional scale measuring Core Depressive Symptomatic Factor and Suicidal and Physical Factor (Shahidi and Shojaee, 2014), and also because the IRT models were not used to analyze its items before, the multidimensional version of GRM (Samejima, 1969) was used to address the research questions in this study. GRM is one of the primaries, natural and well-developed IRT models for polytomous graded responses (Jansen and Roskam, 1986 as cited in Rubio; *et al*., 2007; Muraki, and Carlson, 1993), and it also works well when the items in a test have different levels of discrimination (a_i). The goal of the GRM is to model the relationship between item responses and the latent trait when the ratings include two or more ordered categories using item parameters (Muraki, and Carlson, 1993). GRM can be used in both versions of

unidimensional and multidimensional forms of scales (Ackerman, *et al*., 2003; Bartolucci, Bacci, and Gnaldi, 2012). In this version of GRM, two-parameter logistic IRT model is used that incorporates both a *i* and b*i* parameters; furthermore, the model is often described as the traditional two parameter normal ogive model (Berkeljon, 2012). Since the link between trait level and item response is defined based on the logistic rather than normal ogive cumulative distribution function, it was argued that the normal ogive and logistic functions are approximately concurrent and usually produce similar item characteristic curves using the following equation which refers to the probability of a correct answer (Berkeljon, 2012).

$$P*_{ki}\left(\theta_j\right) = \frac{\exp\left(Da_i\left(\theta_j - b_{ik}\right)\right)}{1 + \exp\left(Da_i\left(\theta_j - b_{ik}\right)\right)}. \qquad (1)$$

In equation 1, k is the ordered response option or score; $P*_{ki}\left(\theta_j\right)$ is the probability of responding to alternative k or above in item i with a trait level θ_j; θ_j is the trait level of the subject; b_{ik} is the location parameter of the alternative k of item i; a_i is the discrimination parameter of item i; and D is the constant (1.7). Based on this equation, in GRM, a-value or item discrimination is a constant value for all category steps in a given item; however, item difficulty or b_i varies across boundaries. These boundaries and their slops were shown in the Figure 1 for item 4. As Figure 1 shows, the item 4 of KADS-11 had a discrimination as a constant value of 1.76 for all boundaries, and item difficulties including b_{i1} = -.095, b_{i2} = 2.36 and b_{i3} = 4.64. Applicable for ordinal options or categories, another function of GRM is to make a distinction between the possible category scores for an item i and the number of steps or boundaries between the category scores (see Figure 1). The possible category scores may vary based on the scoring scale of an instrument such as 0 (no credit), 1 (partial credit), and 2 (full credit) in educational achievement tests or 0 (hardly ever), 1 (much of the time), 2 (most of the time), and 3 (all of the time) in psychological tests. In this model, the more steps successfully completed, the larger the category score with higher scores indicate greater ability or having higher level of measuring trait (Penfield, 2014). Through this model, item characteristic curves (ICC) come from subtracting each of the adjacent ICCs starting with $P_0(\theta)$ using the following formula:

$$P_{ik}\left(\theta_j\right) = P*_{ki}\left(\theta_j\right) - P*_{ki+1}\left(\theta_j\right). \qquad (2)$$

where k: Ordered response option or score; $P_{ki}\left(\theta_j\right)$: Probability of responding to alternative k of item i with a trait level θ_j.

Based on the rationales and suitability of the GRM for KADS-11, the researchers analyzed the item difficulty, item discrimination, test information function, the standard error of measurement and the item information function of the scale through using R software. Additionally, the most psychometric properties such as the construct validity and the reliability of KADS-11 were also analyzed based on CTT through using SPSS to provide a pathway for comparing both IRT and CTT. Using the principal component analysis with the Varimax rotation for the factorial structure of instrument, the analysis was restricted by the following criterion to select the items for a factor: 'an item must have loaded at least 0.40 on its own factor but less than 0.40 on any other factor'. To estimate the reliability of the data, alpha coefficient was computed as a measure of internal consistency.

RESULTS

Construct Validity: Using factor analysis based on the results of principle components, Kaiser-Meyer-Olkin measure of sampling adequacy (Fabrigar, and Wegener, 2012), was completely satisfactory significant (0.91, P < 0.000). After using Varimax rotation, the results revealed that two extracted factors (see Table 2) could explain 55.20% of total variance and proved to be the maximum number interpretable. These factors were (a) Core Depressive Symptomatic Factor (items 1, 2, 3, 4, 5, 6, 7, 8, and 9) and (b) Suicidal and Physical Factor (items 10 and 11). The correlations among factors and total score of KADS-11 were shown in Table 3. The correlation between factor 1 (Core Depressive Symptomatic Factor) and factor 2 (Suicidal and Physical Factor) was .484 (p< 0.01). This value was low enough to show distinct factor with no significant overlap, and it was also satisfactory correlation to maintain the total internal consistency of the KADS-11. In addition, both factors had acceptable correlation with the whole test score, which were statistically significant, indicating high level construct validity for Factor 1, r = .989 (P < 0.000), and for Factor 2, r = .607 (p< 0.01).

Reliability: Calculating alpha coefficient, the internal consistency of the two above factors and the total score of KADS-11 indicated a value of 0.87 for Core Depressive Symptomatic factor and 0.56 for the Suicidal and Physical factor. Coefficient alpha for total items after factor analysis was 0.88. It might be due to the fewer number of items in Factor 2 caused lower internal coefficient. However, it is an adequate and acceptable alpha coefficient based on psychometric assumptions in factor analysis (Beshlideh, 2012). Using the split-half method revealed an acceptable alpha coefficient for internal consistency 0.791 for part 1, 0.763 for part 2 and 0.906 for all items.

KADS-11 item parameters in CTT: Unlike Shahidi's and Shojaee's (2014) study in which item parameters of KADS-11 were not examined, this study focused on KADS-11 item parameters in CTT including item discrimination and item difficulty. To analyze the item discrimination of KADS-11, the correlation of each item with the whole test was calculated (see Table 4). The results revealed that the values were high enough for each item and all of them were highly significant with the range between 0.60 (for item 3) and 0.89 (for item 10). This indicates that KADS-11 items had enough power to distinguish participants who have high depression from those who do not.
In CTT, particularly in achievement or aptitude tests, p-value or '*proportion correct'* is usually considered as an item difficulty; however, this idea is not very meaningful for psychological tests (such as KADS-11), since there is no correct or wrong answer. Alternatively, computing the mean of each item was recommended (Rubio *et al*., 2007) as the item difficulty (see Table 4). Using this method, the results indicated that the extent to which the mean of each item was high, the lower amount of a trait was expected to respond higher categories.

Table 1. General Processes of Test Development

Steps	General Processes of Developing a Test
Step 1	Preparation of test specifications
Step 2	Reviewing the provided item pool generation based on all central components
Step 3	Selecting the first generation of items for the first draft of instrument
Step 4	Reviewing and reduce the items based on the maximum intra-rater validity for each item and ranking the items
Step 5	Designing the most appropriate sampling method to select a pilot group sample
Step 6	Review the sampling method to prevent sampling errors
Step 7	Administering the first version of the scale and pursuing the statistical methods to analyze its primary psychometric properties
Step 8	Technical analyses (e.g., compiling norms, standard setting, equating scores, reliability and validity studies).
Step 9	Determining the last version of the scale based on previous steps.
Step10	Preparation of administrative instructions and technical manual Printing and distribution of tests and manuals.

Table 2. Two-Factor Loadings for Varimax two-Factor Solution

Items	Factor Loadings	
Factor 1:	Core Depressive Symptomatic Factor	
	6. Feeling tired, feeling fatigued, low in energy, hard to get motivated, have to push to get things done, want to rest or lie down a lot.	.75
	7. Trouble concentrating, can't keep your mind on schoolwork or work, daydreaming when you should be working, hard to focus when reading, getting "bored" with work or school.	.72
	9. Feeling worried, nervous, panicky, tense, keyed up, anxious.	.71
	1. Low mood, sadness, feeling blah or down, depressed, just can't be bothered.	.69
	5. Feelings of worthlessness, hopelessness, letting people down, not being a good person.	.68
	4. feeling decreased interest in: hanging out with friends; being with your best friend; being with your boyfriend/girlfriend; going out of the house; doing school work or work; doing hobbies or sports or recreation.	.68
	8. Feeling that life is not very much fun, not feeling good when usually (before getting sick) would feel good, not getting as much pleasure from fun things as usual (before getting sick).	.67
	2. Irritable, losing your temper easily, feeling pissed off, losing it.	.66
	3. Sleep difficulties-different from your usual (over the years before you got sick): trouble falling asleep, lying awake in bed.	.59
Factor 2:	Suicidal and Physical Factor	
	11. Thoughts, plans or actions about suicide or self-harm.	.84
	10. Physical feelings of worry like: headaches, butterflies, nausea, tingling, restlessness, diarrhea, shakes or tremors.	.76

Note: N = 277.

Table 3. Correlations between the Factors of the KADS-11

Factors	Core Depressive	Suicidal and Physical	Total KADS
Core Depressive	1.000		
Suicidal and Physical	.484**	1.000	
Total KADS	.989***	.607**	1.000

Note: N = 277. **p <.01, ***p < .001.

Table 4. Item Parameters in CTT Analysis

Item	Discrimination	Mean Score	SD
1	0.727**	.78	.856
2	0.71**	1.14	.880
3	0.600**	.65	.846
4	0.71**	.70	.829
5	0.759**	.65	.836
6	0.716**	1.01	.887
7	0.725**	1.04	.869
8	0.747**	.78	.870
9	0.726**	.95	.902
10	0.888**	.46	.709
11	0.779**	.20	.520
Total	1.00	8.37	6.11

Note: **p < 0.01 (2-tailed).

Table 5. Estimated Item Parameters of KADS-11

Item	Discrimination for Factor 1	Discrimination for Factor 2	Difficulty 1	Difficulty 2	Difficulty 3
1	1.948	0	-0.329	2.07	4.917
2	1.782	0	-1.606	0.977	3.903
3	1.167	0	0.319	1.862	4.172
4	1.762	0	-0.095	2.36	4.646
5	2.287	0	0.249	2.685	5.601
6	1.773	0	-1.139	1.287	4.167
7	1.812	0	-1.351	1.419	4.141
8	2.171	0	-0.313	2.283	4.983
9	1.763	0	-0.905	1.529	3.951
10	0	1.48	0.831	2.946	5.286
11	0	2.752	3.289	5.838	8.461

Table 6. Comparison of Category Boundaries in CTT and IRT

Item	PC_0 -$\sum PC_{1,2,3}$	d_1 IRT	$PC_{0,1}$ -$\sum PC_{2,3}$	d_2 IRT	$PC_{0,1,2}$ -$\sum PC_3$	d_3 IRT
1	-6.9	-0.329	58.9	2.07	92.9	4.917
2	-48	-1.606	32.8	0.977	87	3.903
3	13.4	0.319	63.2	1.862	94.2	4.172
4	-0.4	-0.095	66.8	2.36	92.8	4.646
5	10.5	0.249	66.1	2.685	93.5	5.601
6	-33.2	-1.139	42.2	1.287	89.2	4.167
7	-40.1	-1.351	43.7	1.419	88.5	4.141
8	-6.8	-0.313	59	2.283	91.4	4.983
9	-26.4	-0.905	48	1.529	87.8	3.951
10	30	0.831	80.6	2.946	97.2	5.286
11	69	3.289	92.2	5.838	98.6	8.461

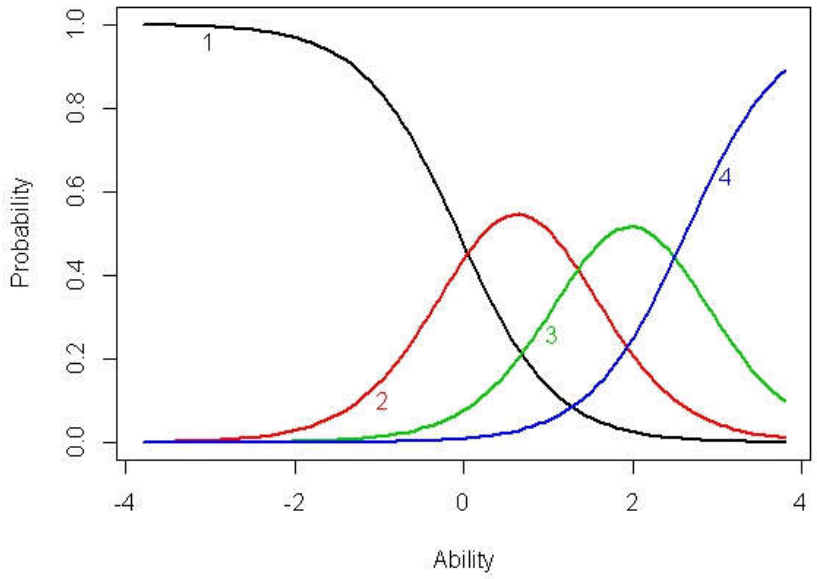

Figure 1. Item category curves of item 4 in the KADS-11

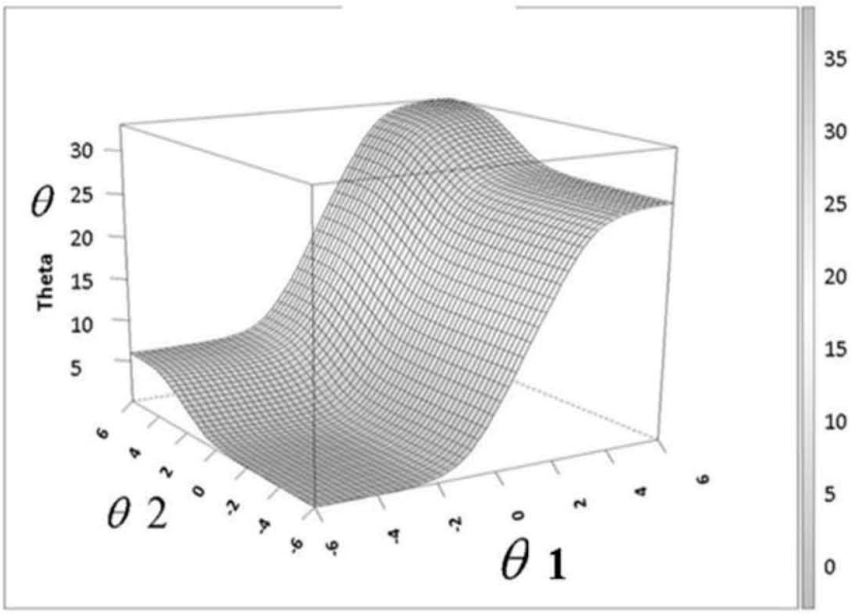

Figure 2. Expected total score diagram and item response surfaces for KADS-11

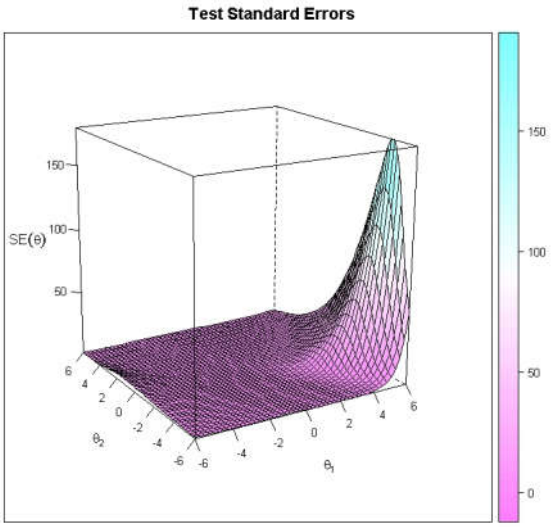

Figure 3. Standard error of measurement (SEM) for KADS-11

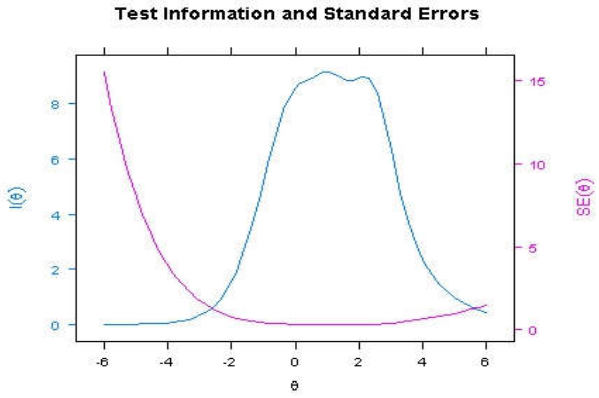

Figure 4. Test information function (the blue line) and standard error of measurement (the red line) in KADS-11

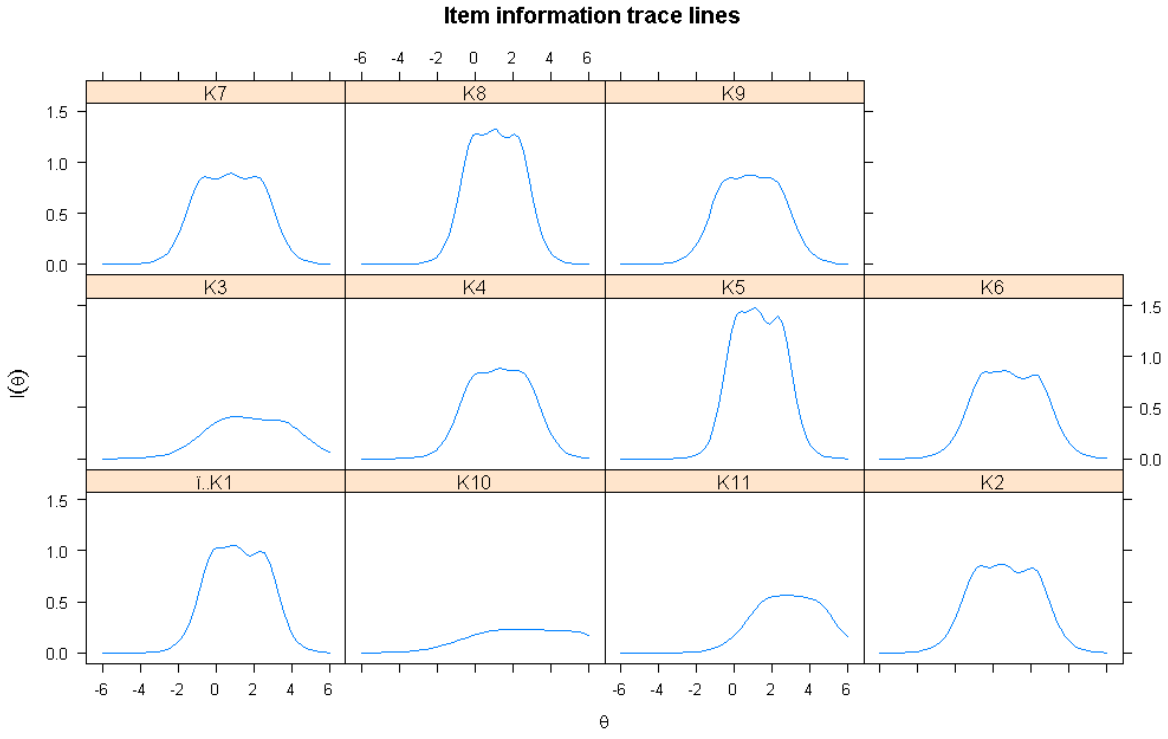

Figure 5. Item information trace line for each item of KADS-11

Although the Table 4 displays this ordinary way of examining the item difficulty, we (the researchers) criticized this method since the method compared with what is usually conducted in IRT is not meaningful enough. Therefore, we replaced it with new method which is argued in the next section.

IRT Analysis of KADS-11

Dimensionality of KADS-11: Since factor analysis indicated two dimensions, factor 1 (Core Depressive Symptomatic Factor) and factor 2 (Suicidal and Physical Factor) for this instrument, the researchers conducted a multidimensional IRT analysis for this data. Analyzing the data with MIRT package (Chalmers, Pritikin, Robitzsch, and Zoltak, 2015) in R (R Core Team, 2014) indicated that the proportion variance explained by the first factor was 43.3% and 10.5% for the second factor and all together both factors could explain 53.8% of the total variability. Also, the correlation between two factors was 0.69 indicating a reasonable association between two factors. According to Tabachnick and Fidell (2007) the correlation between factors should be moderate enough not to have redundancy and in this study the correlation (0.69) was not considered very high to show redundancy and it was not very low to show a completely separation area of variability.

Parameterization: Item parameters of the KADS-11 were shown in Table 5. Regarding a-value (item discrimination), all the items displayed a very rational level of discriminant ranging from 1.167 (for item 3) to 2.287 (for item 5) in the first factor, and 1.48 (for item 10) to 2.752 (for the item 11) in the second factor. To provide a logical way to comparing CTT and IRT parameters for KADS-11, the correlation coefficient between item discriminations in two measurement theories was calculated r = 0.336, which was not a very high correlation. Regarding the item difficulties (b-values) in the psychological instruments, the extreme b_{i0}shows the least amount of the trait needed for choosing first option, and b_{i3} indicates the maximum level of the trait needed to move from pervious category to last category. Since KADS-11 is an instrument measuring depression and our participants were from a normal population, to have a better understanding of the interpretation of b_{ik}, the researchers inversed the signs of the coefficients. For example for item 1, $b_{1,1}$ was 0.329 and $b_{1,3}$ was -4.917 and after reversing we had $b_{1,1}$ = -0.329 and $b_{1,3}$ = 4.917. Based on this inversion, to choose the category 1 a minimum trait level (–0.329) with the probability of 0.5 is required; whereas, approximately a high trait level (4.917) with the probability of 0.5 is needed to choose the category 3; that is the highest level of the trait. Accordingly, individuals with a high level of measuring trait tend to show the maximum depression level in this item. In order to have a meaningful comparison between item difficulties in CTT and IRT, the researchers needed to calculate the b-value differently based on the regular method in CTT. The item difficulty in CTT is the '*proportion correct'* of the items in a sample which is called p-value. Since we don't have the right or wrong answer in psychological tests such as KADS-11, some people estimate the mean of each item for item difficulty (Rubio *et al*., 2007). However, as it was previously noted, to make a comparison between IRT and CTT in the multidimensional GRM, this estimation might not be meaningful enough because there is only one b-value in CTT for item difficulty; whereas, there are three thresholds or boundaries (*b*-values) in IRT for each item with 4-point Likert format.

To solve this problem, the statistical techniques to compute the boundaries in GRM led the researchers to find another novel and meaningful method. In this case, the percentages of responses in each category (PC_0) were calculated; then, the sum of the categories percentages ($\sum PC_{1,2,3}$) was subtracted from the desired category ($PC_0 - \sum PC_{1,2,3}$). The rationale for this new way is that "GRM uses the cumulative segmentation method to estimate the parameters in other words; the k response categories become k-1 dichotomized options" (Rubio, et. al., 2007, p.41). Now, the outcomes were comparable to the GRM boundaries reasonably. For instance, in item 1 the percentages of participants' responses were 46.6%, 32.9%, 17%, and 3.6% respectively for the categories 0 to 3. Based on GRM in which the thresholds are computed cumulatively, the percentage of the first category was subtracted by the sum of the percentages of the other three categories (see the following calculations):

a) For the first boundary: 46.6 – (32.9 + 17 + 3.6) = -6.9
b) For the second boundary: (46.6 + 32.9) – (17 + 3.6) = 58.9
c) For the third boundary: (46.6 + 32.9 + 17) – 3.6 = 92.9

These three values, -6.9, 58.9, and 92.9, are reasonably comparable with the boundaries yielded by GRM in IRT (see Table 6). The correlation coefficients between each boundary in CTT and IRT were also computed resulting 0.971 for the first boundary, 0.899 for the second boundary, and 0.782 for the third boundary. Notably, all correlations are highly positive, which can be a confirmation for the used method in comparison. As previously mentioned, "$PC_0 - \sum PC_{1,2,3}$" comes from the difference between the percentage of a desired category and the sum of the other category's percentages; therefore, the above noted correlation coefficients demonstrate that the extent to which we have the higher "$PC_0 - \sum PC_{1,2,3}$", the higher the trait level is needed to respond to the upper categories of the scale. This conclusion aligns with the interpretation of the boundaries or item difficulties based on GRM.

Model fit: Inspecting whether GRM is an appropriate model for this type of data, a few matters were checked: First, the accuracy of the parameters; item difficulties or boundaries, item discriminations and person parameter or level of the participants' ability were estimated by IRT. Second, expected total score diagram and finally the standard error of measurement were examined. As previously mentioned, the used method in comparing GRM boundaries and the differences between category percentages was supported by the highly positive correlation coefficients, 0.971, 0.899 and 0.782, ($p < .001$). Regarding item discrimination, even though the correlation between a-values in both CTT and IRT was not very high (0.336), it was statistically significant ($p < .01$). In regard to person parameter (ability for participants), theta θ for all 277 participants along with their standard measurement error was computed. The theta values ranged from -1.78 to 2.72 with the standard error of 0.56 for the lowest and 0.35 for

the highest theta in factor 1, and ranged from -1.31 to 2.62 with standard error of 0.56 for the lowest and 0.77 for the highest theta in factor 2. Statistically, this range represents a normal distribution of the depression trait in our normal sample population. Thus, the person parameters supported this idea that most of the participants are in the mid-level of the trait (depression) distribution, also the level of the standard errors were reasonable. Additionally, since two measuring factors are correlated to each other (r = .691), this multidimensional data are categorized in compensatory model. The model indicates that if an examinee has high ability in one dimension, this high ability may compensate his or her low ability in other dimensions. In other words, having high level of one factor can cause to be high in the other factor too. Supporting this idea, the '*expected total score diagram*' was depicted, and it displayed that people with low level of factor 1 (Core Depressive Symptomatic Factor) tend to choose the options (categories) with the low loading of factor 2 (Suicidal and Physical Factor). Similarly, one with high level of factor 1 has also high level of factor 2 (see Figure 2).

The last evidence for fitting GRM to this data is the Standard Error of Measurements (SEM). The diagram of '*Test Standard Error*' displayed that when there are not enough responses in a specific level of *theta*, the level of error goes up. That is, there are very small amount of error in estimating the low and moderate level of depression, and there are considerable amount of error to estimate the higher degree of depression (see Figure 3). This result aligns with the nature of the sample population in this study within which there were not clinical diagnosed depressed individuals. However, since there are not many high depressed individuals in the sample population, there are not enough responses at the other specific levels of ability. Therefore, the precision of estimation tend to be low (De Champlain, 2010). Moreover, comparing the plot of *Test Information* and simple SEM plot supports the abovementioned idea. This plot showed that most gained information come from middle level of the trait, which is between -3 and +3, and the lowest level of measurement error occur in this interval (see Figure 4).

DISCUSSION

One of the advantages in analyzing the data with IRT is that the researchers could have more technical information about each item of a test. This valuable information led to optimize developing KADS-11. In this study to visualize the amount of information that can be yielded from each item was depicted in the *'Item Information Trace Line'* (see Figure 5). As shown in the Figure 5, most of the items provide the maximum amount of information close to 1.00, which implies a reasonable amount of information about the latent trait (depression) in the participants. Particularly, items 8 and 5 gained the most amount of information indicating that these items are key items in KADS-11. The other notably fact was that most of the items were centralized on the middle range of the trait, and there was not much information about the two extreme levels of the trait. This characteristic is compatible with the nature of sample population which was a normal population. However, the items 3, 10 and 11 could provide some information about the higher level of the depression. In regard to parameters and their comparison in two measurement theories, the item discriminations in both theories were computed. The correlation between these item discriminations was not very high (r = .336) but significant (P < .01). Scrutinizing the reason, it was revealed that the item 10 reduced the expected correlation coefficient. This correlation without item 10 improved significantly (r = 0.91). Item 10 showed a high level of discrimination power in CTT, but slightly low coefficient in IRT analysis. Particularly, examining the item information functions revealed that this item does not provide enough information; thus, revising this item is highly recommended to improve the instrument. Comparing item difficulties in CTT and IRT revealed that there are highly positive correlations between item difficulties. Since in GRM the boundaries are considered as the item difficulty in which the number of boundaries is k-1, and because in CTT there is just one value for item difficulty, the researchers attempted to make a meaningful comparison between these values by using the difference between the percentage of a desired category and the sum of the other category's percentages. This new statistical technique allowed the researchers to be able to compare GRM boundaries with item difficulties in CTT (see Table 6). Generally, in regard to the comparability of the parameters in this research, it is deduced that parameters in CTT and IRT are comparable to each other.

Another important fact that is worth to mention is about the standard error of measurement (SEM). In CTT standard error is assumed to be a constant value across all items and examinees; whereas, in IRT it varies through the different level of the trait. Scrutinizing the SEM in different level of the trait in the diagram of '*Test Standard Error*' displayed that the level of error goes up when there are not enough responses in that specific level (see Figure 3). Specifically, in our sample, which was derived from a normal population, there were a few depressed people with high level of the latent trait. Therefore, the standard error was high in this level of the trait. Similarly, the level of error went down in the middle and low level of the trait in the diagram that showed the number of undepressed examinees was high in the sample population. In this regard, it can be concluded that IRT gives a better understanding of the standard error, and can provide researchers with a suitable guide to control such an undesired factor. In conclusion, since the parameterization is comparable in both measurement theories, and because IRT has many advantages in analyzing this data including extra technical information about each item, better treatment of the SEM and better model fit, it can be deduced that the IRT provides the researchers with a more useful and comprehensive information regarding the multidimensional instruments and the participants simultaneously.

Limitation: Although the scale was analyzed based on Iranian sample, the instrument can be utilized in other Iranian groups who are living in other countries (e.g., newcomers in Canada) based on the main common characteristics of individuals. However, it is suggested to research the psychometric properties of the scale in other ethnical groups. Thus, we do not generalize the results to other ethnical groups. With regards to the dimensions of this scale, it is suggested that the two items in Suicidal and Physical Factor are not enough to

measure individuals' suicidal thoughts and the physical symptoms of depression. Although, this research showed that these items (e.g., 10 and 11) determined significantly a unique factor, these items should be increased to a reliable level in the next new version of scale.

REFERENCES

Ackerman, T. A., Gierl, M. J. and Walker, C. M. 2003. Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*, 37-53.

Adler, M., Hetta, J., Isacsson, G. and Brodin, U. 2012. An item response theory evaluation of three depression assessment instruments in a clinical sample. *BMC Medical Research Methodology,* 12-84.

Allqaier, A.K., Pietsch, K., Fruhe, B., Sigl-Glockner, J. and Schulte-Kome, G. 2012. Screening for depression in adolescents: validity of the patient health questionnaire in pediatric care. *Depress Anxiety*, 906-913. doi: 10.1002/da. 21971.

Barthel, D., Barkmann, C., Ehrhardt, S. and Bindt, C. 2014. Psychometric properties of the 7-item generalized anxiety disorder scale in antepartum women from Ghana and Cōte d'Ivoire. *Journal of Affective Disorders,* 169, 203–211.

Bartolucci, F., Bacci, S. and Gnaldi, M. 2012. MultiLCIRT: An R package for multidimensional latent class item response models. arXiv:1210.5267v1 [stat.AP].

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. L., Kendall, D. and Yorkstona, K. 2011. An introduction to item response theory and rasch models for speech- language pathologists. *American Journal of Speech-Language Pathology,* 20, 243–259.

Bech, P., Paykel, E., Sireling, L. and Yiend, J. 2015. Rating scales in general practice depression:Psychometric analyses of the Clinical Interview for Depression and the Hamilton Rating Scale. *Journal of Affective Disorders*, 171, 68–73.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J. and Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry,* 4,561–71.

Berkeljon, A. 2012. Multidimensional item response theory in clinical measurement: A bifactor graded response model analysis of the outcome-questionnaire-45.2 (Brigham Young University). *ProQuest Dissertations and Theses,* Retrieved fromhttp://scholarsarchive.byu.edu/cgi/view content.cgi?article=4567andcontext=etd on Apr30th, 2015.

Beshlideh, K. 2012. *Research Methods and Statistical Analysis of Research Examples Using SPSS and AMOS.* Shahid Chamran University Inc. Ahvaz, Iran.

Bottesi, G., Ghisi, M., Altoè, G., Conforti, E., Melli, G. and Sica, C. 2015. The Italian version of the depression anxiety stress scales-21: Factor structure and psychometric properties on community and clinical samples. *Comprehensive Psychiatry, Elsevier Inc*, 60, 170–181.

Broen, M. P. G., Moonen, A. J. H, Kuijf, M. L., Dujardin, K., Marsh, L., Richard, I.H, Starkstein, S. E., Martineze Martin, P. and Leentjens, A. F. G. 2015. Factor analysis of the Hamilton Depression Rating Scale in Parkinson's disease. *Parkinsonism and Related Disorders, 21,* 142-146.

Brooks, S. J. and Kutcher, S. 2001. Diagnosis and measurement of adolescent depression: a review of commonly utilized instruments. *Journal of Child and Adolescent Psychopharmacology, 11*(4), 341-376. doi:10. 1089/104454601317261546.

Brooks, S. J., Krulewicz, S. P. and Kutcher, S. 2003. The Kutcher adolescent depression scale: assessment of its evaluative properties over the course of an 8-week pediatric pharmacotherapy trail. *Journal of Child and Adolescent Psychopharmacology*, *13*(3), 337-349.

Bulut, O. 2014, April. Examining subscore reliability within the multidimensional IRT framework. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Canadian Mental Health Association (CMHA). 2013. *Fast Facts about Mental Illness*. http://www.cmha.ca/media/ fast-facts-about-mental-illness/#.VdPGlPlViko

Carneiro, A. M., Fernandes, F. and Moreno, A. R. 2015. Hamilton depression rating scale and montgomery–asberg depression rating scale in depressed and bipolar I patients: psychometric properties in a Brazilian sample. *Health and Quality of Life Outcomes*, 13-42. DOI 10.1186/s12955-015-0235-3

Carroll, B. J., Fielding, J. M. and Blashki, T. G. 1973. Depression rating scales: a critical review. *Archives of General Psychiatry, 28,*361-366.

Centre for Addiction and Mental Health (CAMH). 2011. *Mental Health Facts and Statistics.* http://wptheme.Cameron helps.ca/wp-content/uploads/2011/12/Mental-Health-Statistics.pdf

Centre for Addiction and Mental Health (CAMH). 2015. *Mental Health Facts and Statistics.*http://www.camh.ca/en/ hospital/about_camh/newsroom/for_reporters/Pages/addictionmentalhealthstatistics.aspx

Chalmers, P., Pritikin, J, Robitzsch, A. and Zoltak, M. 2015. mirt: multidimensional item response theory. [Computer software]. Available from http://CRAN.Rproject. org/package=mirt.

Cronbach, J. L. and Meehl, E. P. 1955. Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Costa, P. T. and McCrae, R. R. 1985. The NEO personality inventory manual. Odessa, FL: Psychological Assessment Resources 1994. Set like plaster? Evidence for the stability of adult personality, in T. Heatherton and J.Weinberger (eds.), *Can personality Change* ?, 21–40.Washington, DC: American Psychological Association.

Costa, P. T., Terracciano, A. and McCrae, R. R. 2001. Gender differences in personality traits across cultures: robust and surprising findings, *Journal of Personality and Social Psychology,* 81, 322–31.

De Champlain, A. F. 2010. A primer on classical test theory and item response theory for assessments in medical education. *Medical Education,* (44) 109–117. doi:10.1111/ j.1365-2923.2009.03425.x

Dere, L., Watters, C. A., Chee-Min Yu, S., Bagby, R. M., Ryder, A. G. and Harkness, K. L. 2015. Cross-cultural examination of measurement invariance of the Beck Depression Inventory–II. *Psychological Assessment,* 27(1), 68–81.

Fabrigar, L. R. and Wegener, D. T. 2012. *Exploratory Factor Analysis*. Oxford University Press, USA.

Hambleton, R. K. and Jones, R. W. 1993. An NCME instructional module on. *Educational Measurement: Issues*

and Practice, 12, 38–47. doi: 10.1111/j.1745- 3992.1993. tb00543.x

Hamilton, M. 1960. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

Heatherton, T.F. and Polivy, J. 1991. Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology*, 60, 895-910.

Hooman, H. A. 2007. *Handbook of qualitative research*. Samt Publication. Tehran, Iran. http://samt.ac.ir

Hooman, H. A. 2010. *Research methodology in behavioral sciences*. Samt Publication, Tehran, Iran. http://samt.ac.ir

Ialomiteanu, A.R., Hamilton, H.A., Adlaf, E.M. and Mann, R.E. 2014. CAMH Monitore Report: Substance Use, Mental Health and Well-Being among Ontario Adults, 1977–2013 (*CAMH Research Document Series, 40*). Toronto, ON: Centre for Addiction and Mental Health. Available at: http://www.camh.ca/en/research/news_and_ publications/Pages/camh_monitor.aspx

Ghanei, R., Golkar, F. and Aminpour, E. 2014. The relationship between depression and self-care in patients with type II diabetes, *Nursing Practice Today,* 1(1), 2-8.

Guerra, M., Ferri, C., Llibre, J., Prina, AM. and Prince, M. (2015). Psychometric properties of EURO-D, a geriatric depression scale: a cross-cultural validation study. *BMC Psychiatry,* 15,12-33. DOI 10.1186/s12888-015-0390-4

Güler, N., Uyanık, G. K. and Teker, G. T. 2014. Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education,* 2(1), 1-6.

Kane, T. M. 1992. An argument-based approach to validity. *Psychological Bulletin,* 112(3), 527-535.

Kerlinger, F. N. 1986. *Foundations of Behavioral Research*, 2nd ed., New York: Holt, Rinehart and Winston, Inc.

Lahlou-Laforêt, K., Ledru, F., Niarra, R. and Consoli, S. M. 2015. Validity of Beck Depression Inventory for the assessment of depressive mood in chronic heart failure patients. *Journal of Affective Disorders,* 184, 256-260.

Lay, C. 1986. At last, my research article on procrastination. *Journal of Research in Personality*, 20, 474-495.

LeBlanc, J. C., Almudevar, A., Brooks, S. J. and Kutcher S. 2002. Screening for adolescent depression: Comparison of the Kutcher adolescent depression scale with the Beck depression inventory. *Journal of Child and Adolescent Pscychopharmacology,* 12(2), 113-126.

Levine, S. Z. 2013. Evaluating the seven-item Center for Epidemiologic Studies Depression Scale short-form: a longitudinal US community study. *Social Psychiatry and Psychiatric Epidemiology,* 48, 1519–1526. DOI 10.1007/ s00127-012-0650-2.

Makaremi, A. 1992. Sex differences in depression of Iranian adolescents. *Psychological Reports,* 71(3), 939-43.

Mead, A. D. and Meade, A. W. 2010. Test construction using CTT and IRT with unrepresentative samples. *Educational and Psychological Measurement.* http://mypages.iit.edu/ ~mead/Mead_and_Meade-v10.pdf

Montgomery, S. A. and Asberg, M. 1979. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry,* 134,382–389.

Muraki, E. and Carlson, E. J. 1993. Full-information factor analysis for polytomous item responses. Paper presented at the annual meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).

Neal, D. J. and Carey, K. B. 2005. A follow-up psychometric analysis of the self-regulation questionnaire. *Psychology of Addictive Behaviors, 19*(4), 414–422.

Neal, D. J. and Carey, K. B. 2004. Developing discrepancy within self-regulation theory: Use of personalized normative feedback and personal strivings with heavy-drinking college students. *Addictive Behaviors,* 29, 281–297.

Penfield, D. R. 2014. An NCME instructional module on polytomous item response theory models, *Educational Measurement: Issues and Practice,* 33(1), 36–48.

Popper, K. R. 2005. *The Logic of Scientific Discovery*. Routledge, London.

Popper, K.R. 1994. *The Myth of the Framework: In the Defense of Science and the Rationality.* Routledge, London.

R Core Team, 2012. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reilly, T. J., MacGillivray, S. A., Reid, I. C. and Cameron, I. M. 2015. Psychometric properties of the 16-item Quick Inventory of Depressive Symptomatology: A systematic review and meta-analysis. *Journal of Psychiatric Research,* 60, 132-140.

Roberge, P., Dore, I., Menear, M., Chartrand, E., Ciampi, A., Duhoux, A., and Fournier, L. 2013. A psychometric evaluation of the French Canadian version of the Hospital Anxiety and Depression Scale in a large primary care population. *Journal of Affective Disorders,* 147, 171-179.

Rubio, V., Aguado, D., Hontangas, P. M. and Hernjndez, J. M. 2007. Psychometric properties of an emotional adjustment measure: An application of the graded response model. *European Journal of Psychological Assessment, 23*(1), 39–46. DOI 10.1027/1015- 5759.23.1.39.

Rush, A. J., Giles, D.E., Schlesser, M.A., Fulton, C.L., Weissenburger, J. and Burns, C. 1986. The inventory for depressive symptomatology (IDS): Preliminary findings. *Psychiatry Research, 18,*65–87.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R., Thase, M. E., Kocsis, J. H., and Keller, M. B. 2003. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry, 54,* 573-583.

Qualter, P., Brown, S. L., Munn, P., and Rotenberg, K. J. (2010). Childhood loneliness as a predictor of adolescent depressive symptoms: An 8-year longitudinal study. *European Child and Adolescent Psychiatry, 19*(6), 493-501. doi:10.1007/s00787-009-0059-y.

Samejima, F. 1969. Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, 34*(17).

Sebille, V., Hardouin, J. B., Le Néel, T., Kubis, G., Boyer, F., Guillemin, F. and Falissard, B. 2010. Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients - a simulation study. *BMC Medical Research*

*Methodology,*10-24. http://www.biomedcentral.com/1471-2288/10/24.

Schellingerhout, M. J., Heymans, W. M., Verhagen, P. A., de Vet, C. H., Koes, W. B, and Terwee, B. C. 2011. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. *BMC Medical Research Methodology,* 11-87. http://www.biomedcentral.com/1471-2288/11/87

Shahidi, M. and Shojaee, M. 2014. Psychometric properties and diagnostic utility of the 11-item Kutcher adolescent depression scale (KADS) in Persian samples. *International Journal of Psychology and Behavioral Sciences, 4*(6), 201-207. DOI: 10.5923/j.ijpbs.20140406.03

Sheng. Y. and Wikle, K. C. 2007. Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 67*(6), 899-919.

Shojaeizadeh, D., and Rasafiyani H.R. 2001. A study on depression among pre-university students Kazeron city. *Rehabilitation, 2,* 68-69.

Steptoe, A., Tsuda, A., Tanaka, Y. and Wardle, J. 2007. Depressive symptoms, socio-economic background, sense of control, and cultural factors in university students from 23 countries. *International Journal of Behavioral Medicine, 14*(2), 97–107. doi:10.1007/BF03004175.

Szumilas, M., Kutcher, S., LeBlanc, J. C., and Langille, D. B. 2010. Use of school-based health centres for mental health support in Cape Breton, Nova Scotia. *The Canadian Journal of Psychiatry, 55*(5), 319-328.

Trujols, J., Feliu-Soler, A., Diego-Adelińo, J., Portella, M. J., Cebrią, Q., Soler, J., Puigdemont, D., Įlvarez, E.,and Perez, V. 2013. A psychometric analysis of the clinically useful depression outcome scale (CUDOS) in Spanish parents. *Journal of Affective Disorders, 151,* 920–923.

Tabachnick, B. G., and Fidell, L. S. 2007. *Using Multivariate Statistics (5th Ed.)* Boston: Allyn and Bacon.

Walkiewicz, M., Tartas, M., Majkowicz, M. and Budzinski, W. 2012. Academic achievement, depression and anxiety during medical education predict the styles of success in a medical career: A 10-year longitudinal study. *Medical Teacher*, *34,* 611–619. doi: 10.3109/0142159X.2012. 687478

World Health Organization, 2015. *Mental Health Atlas*. http://www.who.int/mental_health/evidence/atlas/mental_health_atlas_2014/en/

Zimmerman, M., Martinez, J. H., Young, D., Chelminski, I. and Dalrymple, K. 2013. Severity classification on the Hamilton depression rating scale. *Journal of Affective Disorders*, *150,* 384–388.

Zoghi, M. and Valipour, V. 2014. A comparative study of classical test theory and item response theory in estimating test item parameters in a linguistics test. *Indian Journal of Fundamental and Applied Life Sciences, 4*(S4), 424-435.

Zung, W.W. 1965. A self-rating depression scale. *Archives of General Psychiatry,* 12,63-70.

****