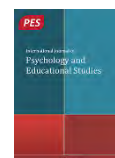





www.ijpes.com



Item Response Theory-Based Psychometric Investigation of SWLS for University Students

Akif AVCU¹
¹Marmara University, İstanbul, Turkey  0000-0003-1977-7592

ARTICLE INFO

Article History:

Received 02.12.2020

Received in revised form

15.02.2021

Accepted 26.03.2021

Available online

21.04.2021

ABSTRACT

Life satisfaction is an important factor for mental health and has many positive effects on people. Considering its importance, different measurement tools were developed over the last 5 years. Among these tools, the Satisfaction with Life Scale (SWLS) is the most prominent and adapted to diverse populations, including university students. On the other hand, all these studies were conducted using the classical testing approach, while item response theory was rarely preferred. Regarding this gap, this study aimed to evaluate the psychometric properties of the Turkish version of SWLS by using a graded response model (GRM), which is a member of a broader family of the modern psychometric approach called Item Response Theory (IRT). For this purpose, the data were collected from 471 university students (male = 83, female = 388) aged between 17 and 37 years ($M = 21.23$, $SD = 2.32$). IRT based analysis provided satisfactory results on the psychometric properties of the SWLS. It was found that the scale's reliability was acceptable across a wide range of ability spectrums and items fit the GRM well and did not show gender-based differential item functioning. As a result, the psychometric quality of SWLS was further proved in the IRT context.

© 2021 IJPES. All rights reserved

Keywords:

item response theory, satisfaction with life, differential item functioning

1. Introduction

In the modern psychology literature, psychopathology-oriented views and studies have been replaced with the ones that focused mainly on positive aspects of human beings. This new branch is called positive psychology, and the researchers' growing interest has reinforced its place in psychology (Masten, 2001). Positive psychology focuses on strengths and is interested in developing these aspects that acts as a buffer against psychopathological problems (Veenhoven, 1988). As positive psychology comes to the fore, life satisfaction concepts have become more popular topics in psychology literature.

Life satisfaction is a determinative factor on individuals' mental health (Diener & Suh, 1997). Pavot and Diener (1993) described life satisfaction as a critical process whereby individuals question the extent to which their life cycle meets their expectations based on their subjective criteria. Life satisfaction of individuals is primarily related to the level of well-being in the society, quality of health services and educational opportunities (Diener & Seligman, 2004). Appleton and Song (2008) suggest that life satisfaction has six different components: (1) income level of the person, (2) occupation and social status, (3) opportunities and social mobility, (4) welfare, (5) existing state policies and, and (6) environment, family and social relations.

Life satisfaction is also closely related to subjective well-being and regarded as its cognitive component (Dorahy et al., 2000). A higher level of life satisfaction has many positive effects on people. Individuals who

¹ Corresponding author's address: Department of Educational Sciences, Marmara University, İstanbul, Turkey

e-mail: avcuakif@gmail.com

Citation: Avcu, A. (2021). Item response theory-based psychometric investigation of SWLS for university students. *International Journal of Psychology and Educational Studies*, 8(2), 27-37

<https://dx.doi.org/10.52380/ijpes.2021.8.2.265>

have positive emotions and perceptions of their lives have more effective problem-solving skills and are more resilient to stressful life events (Matheny et al., 2002). Additionally, life satisfaction is seen as a determinative of life quality (Iwasa et al., 2006). In general, individuals with higher life satisfaction are more compatible with and productive in society.

There is an increase in the number of studies dealing with different aspects of life satisfaction in the last few decades, considering the importance of this construct (Baird et al., 2010). valid and reliable measurement tools were needed during the data collection process. As a result of this growing interest, many different scales have been developed to measure life satisfaction to increase the validity of studies: The Satisfaction with Life Scale (SWLS: Diener et al., 1985); The Brief Life Satisfaction Scales (Lubin & Van Whitlock, 2004); The Temporal Satisfaction with Life Scale (Pavot et al., 1998); The Riverside Life Satisfaction Scale (Margolis et al., 2018); Student's Life Satisfaction Scale (Huebner, 1991); and Life Satisfaction Measure based on Judgment Theory (Meadow., 1992).

Among them, the SWLS is the most prominent one. It has already been adapted to various populations worldwide: Brazilian university students (Zanon et al., 2014); Turkish university students and elderly adults (Durak et al., 2010); Hong Kong university students (Sachs, 2003); Spanish adults (Vázquez et al., 2013); Mexican adults (López-Ortega et al., 2016); and people with Parkinson's disease (Lucas-Carrasco et al., 2014) etc. All these studies have unanimously reported good psychometrical properties of the SWLS.

On the other hand, including the first development study of SLWS (Diener et al., 1985), more conventional statistical methods were preferred in all of these studies to unveil psychometrical characteristics of SLWS. These methods belong to the traditional test development approach called Classical Test Theory (CTT). The most important advantage of the classical approach is the familiarity of the concepts by the researchers.

According to Embretson (1996), studies based on the CTT approach have some limitations. First, it is assumed that reliability in the CTT approach is usually fixed for all possible ability scores (ability is generic term to refer to the trait being measured). Secondly, measurement characteristics are considered to be related to the sample group. Therefore, for each study carried out with different samples, validity and reliability need to be investigated again. On the other hand, the IRT scales both items and person's latent trait level on the same metric and provide sample free statistics: item statistics are independent of the sample group and person-level statistics are independent of a specific set of items (Hambleton et al., 1991). Thanks to this approach's superiority, IRT's use has become widespread in test development, validity and reliability studies.

IRT was first noted in the 1970s when it was used to develop standardised tests, such as Scholastic Aptitude Tests-SAT. It has become a widely used psychometric method in the validity studies of measurement tools (Samejima, 1969). The IRT is based on the idea that a person's response to test items depends on only two factors: the person's ability level (denoted as θ) and the item's characteristics (Bond & Fox, 2001). Regarding the superiority of IRT, it is assumed that the psychometric evaluation of SLWS will enrich our insights on the construct of life satisfaction and further support the well-proven psychometric quality of SLWS in this new and promising framework. Based on this reality, the current study aimed to evaluate the psychometric properties of SLWS with IRT for university students.

1.1. Overview of IRT

In IRT literature, many different models characterise the item properties. The earlier models were proposed to model dichotomously scored achievement tests. One, two and three parameter logistic models were proposed for such tests (Birnbaum, 1968). Simultaneously, the models belonging to the Rasch analysis family were also developed (Rasch, 1960). In the following years, the two-parameter logistic model has been generalised to scale items with multiple response categories, resulting in polytomous response models. Among these models, the most popular ones are the Graded Response Model (Samejima, 1969), Partial Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1992; 1993), and the Rating Scale Model (Andrich, 1978a; 1978b). When conducting IRT based estimation, model fit indices of these models are compared, and the best fitting model is selected to estimate the data. These models provide similar outputs and when one model fit the data well, the other usually fit well (Maydeu-Olivares et al., 1994). On the other hand, for the items with Likert type ordered response categories, Graded Response Model (GRM) is usually preferred.

In the GRM models, the item's properties are depicted by the discrimination parameters (a) and the difficulty parameters and (d) by using category response functions. The difficulty parameter is also referred to as the location or intersection parameter. The precision of these estimated parameters is also indicative of information obtained for the respondents. Statistically, how precisely the parameters are estimated is related to the variance of the predicted parameters. The variance of these values is expressed as σ^2 . Information is expressed with "I" and the amount of information is determined as $1/\sigma^2$. From a psychometric perspective, information indicates how much a parameter is predicted.

The amount of information is also closely related to the discrimination parameters: the higher the the item's discrimination parameter, the more information it provides. The amount of information is calculated at the item level. By assessing the amount of information of each item, good working items could be specified. Additionally, item information could be depicted with item information functions (IIFs). Having pre-knowledge of the amount of item information, better measurement tools could be constructed. The information functions at the test level can also be obtained by summing up the item information functions. Item information functions and test information functions enable graphical investigation of tests. Those graphs show which ability levels items and tests provide most information (and more precise estimation). As underlined above, the amount of information is directly related to the a parameter. The higher the discrimination parameter, the greater the amount of information item provides. The item's difficulty level also indicates ability spectrum's the location where the maximum information can be obtained (Hambleton et al., 1991).

Theoretically, GRM is an extension of the two-parameter logistic model (Birnbaum, 1968). In the GRM framework, the number of dichotomies corresponds to the number of categories minus one. For example, there are four different dichotomies for a 5-point Likert-type item. These dichotomies are progressively compared: category 1 is compared with 2, 3, 4 and 5; categories 1 and 2 are compared with 3, 4 and 5; categories 1, 2 and 3 are compared with 4 and 5 and categories 1,2,3 and 4 are compared with category 5. The category boundary response function (CBRF) is calculated for each dichotomy using the generalised 2PLM. They are also known as item trace lines. Finally, the item characteristic response functions (ICRF) values are calculated using the obtained CBRFs which depicts the probability of responding to any response options' category. The respondent's performance and the characteristics that underlie item performance can be described by these monotonically increasing functions (Henard, 2000). ICRFs are usually sigmoid curves that describe a change in the likelihood of the response based on the individual's latent trait level. The shape of the ICRFs is determined by the item characteristics predicted in the model. Item characteristic curves provide important information about the properties of items. These curves graphically show the likelihood of item responses throughout different ability levels. As a result of examining the graphs, it is possible to determine weak and overlapping item categories.

IRT-based models have three basic assumptions about the data: (a) unidimensionality, (b) local independence, and (c) the IRT model fitting to the data (Reeve & Fayes, 2005). Unidimensionality means that there is only one factor affecting an individual's performance on an item. That is, the model has a single θ value for each individual. Violating this assumption will result in an incorrect estimation of the standard error and parameters (DeMars, 2010). Unidimensionality could be tested with exploratory factor analysis and confirmatory factor analysis. When using exploratory factor analysis, the first factor's amount of variance is evaluated for possible unidimensionality. If the first dimension explains more than 20% of the variance, an underlying dominant factor's existence can be justified (Hattie, 1985). Additionally, confirmatory factor analysis could be performed to test one dimensional model. Also, the violation of unidimensionality can be caused by the differentiation of the item properties across demographic groups. When this occurs, the response patterns become a function of both the underlying trait and the group membership. This problem is difficult to diagnose with a factor analytic approach. There are many methods available for determining whether items show differential item functioning (DIF). For further reading on these methods, please see Holland and Wainer (1993). Among these methods, the logistic regression method was used in the current study because it allows detecting both uniform and non-uniform DIF for polytomous data (Swaminathan & Rogers, 1990). This method is based on the likelihood ratio χ^2 statistics. With this approach, three hierarchical nested models are created and DIF is classified as uniform or nonuniform and total. Uniform DIF can be detected by comparing Model 1 and Model 2, non-uniform DIF can be detected by comparing Model 2 and Model 3 and

total DIF can be detected by comparing Model 1 and Model 3. Statistics. Pseudo R^2 statistics could evaluate the magnitude of the DIF. Zumbo (1999) stated that the following threshold values of R^2 could be used for evaluation of DIF: negligible DIF (<0.13), moderate DIF (between $0.13 - 0.26$), and large DIF (> 0.26).

Local independence indicates that there is no relationship between item responses when θ is kept constant. Specifically, for any ability level, the probability of responding to an item is independent of the probability of the response to any other items. For the current study, the Q3 statistics proposed by Yen (1984) were used to test the local dependence because it is similar to and interpreted as the same with well-known Pearson correlations. If Q3 is found to be 1, the items are considered perfectly dependent, and 0 means perfect independence.

2. Method

2.1. Participants

The current study data were selected via an online data collection tool due to the 2020 pandemic outbreak. Because it is not easy to apply probability-based sampling techniques when using online platforms, the convenience sampling technique was preferred to collect the data. In the sample group, there are 471 university students (388 females and 83 males). Students were enrolled in 4 different faculties (Faculty of Science and Literature, Faculty of Economics and Administrative Sciences, Faculty of Health Sciences and Faculty of Fine Arts). The participants' ages ranged between 17 and 37 years ($M = 21.23$, $SD = 2.32$).

2.2. Measurement tools

The SWLS, was used to collect the data. It was developed by Diener et al., (1985). The SWLS was adapted to the Turkish population by Köker (1991) and the re-evaluation of the scale's validity was performed by Dağlı and Baysal (2016). They found that the internal consistency coefficient of SWLS was 0.88 and test-retest reliability was 0.97. Additionally, both EFA and CFA results suggested that the SWLS is a unidimensional scale, and the dimensional structure was compatible with the original scale. The scale consisted of five Likert type questions. Higher scores imply a higher level of life satisfaction levels.

2.3. Analysis

As the first step of the data analysis process, assumptions of IRT were tested. Unidimensionality was assessed with both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). SPSS (IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.) was used to perform EFA and MLUS 6 (Muthén & Muthén, 1998-2011) was used to perform CFA. The remaining analyses were performed in R statistical environment (R Core Team, 2017). Further, gender based DIF was assessed using the "lordif" package, developed by Seung, Gibbons & Crane (2016). Finally, the "mirt" package, developed by Chalmers (2012), was used to conduct IRT-based model estimation, and evaluate the local independence assumption. The model fit statistics, item-fit statistics (Kang & Chen, 2007), item parameters, the amount of information each item yields, test information statistics, and test reliability function were estimated during the model estimation process.

3. Results

3.1. Checking the IRT assumptions

This study aims to examine the psychometric properties of the SWLS using GRM and to obtain further evidence for its validity by performing analysis in the IRT context. Before starting, IRT assumptions (unidimensionality and local independence) were tested. Testing of the unidimensionality was carried out in two steps.

In the first step, explanatory factor analysis was conducted. For this analysis, the amount of variance explained by the first factor and factor loadings of each item were investigated. It was revealed that the variance explained by the first factor was 56% and the factor loadings of the items ranged from 0.71 to 0.80. These findings supported that there is a dominant underlying factor explaining the variance of the SWLS items. Also, Cronbach's alpha coefficient was estimated as 0.80. It showed that the internal consistency of the items was sufficiently high. In the second step, one dimensional CFA was fitted to the data to test the unidimensionality

of the SWLS. The findings showed that the unidimensional structure of SWLS was confirmed [$\chi^2(5) = 10.765$, $\chi^2/df = 2.15$, CFI = 0.99, TLI = 0.98, RMSEA = 0.049 (95% CI = 0.001-0.091), SRMR = 0.018].

The unidimensionality assumption was further investigated by investigating DIF. Lord's chi-square (χ^2) test statistic was calculated to determine whether the items showed differential item functioning based on the gender variable. The results were given in Table 1 below. The table contains the statistics obtained by comparing three models (Model 1 vs. Model 2 for uniform DIF, model 1 vs. model 3 for non-uniform DIF and model 2 vs. model 3 for total DIF). The pseudo R^2 values (Cox, & Snell, 1989; Nagelkerke, 1991) showed no statistically significant χ^2 differences $p < 0.05$ level. These findings indicate that no item showed DIF based on the gender variable and can be considered additional proof for the scale's unidimensionality. Additionally, DIF is also related to the validity of the scale. Therefore, findings related to DIF can be regarded as supporting evidence to the validity of the scale.

Table 1. Lord's Chi Square Test Statistics to Investigate Differential Item Functioning of SWLS

χ^2 values			Peudo R^2 Values					
			Nagelkerke			CoxSnell		
χ^2 1-2	χ^2 1-3	χ^2 2-3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
0.663	0.898	0.874	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000
0.365	0.345	0.253	0.001	0.002	0.002	0.001	0.002	0.001
0.081	0.097	0.204	0.002	0.003	0.001	0.002	0.003	0.001
0.078	0.049	0.086	0.003	0.006	0.003	0.003	0.006	0.003
0.429	0.091	0.041	0.001	0.005	0.005	0.001	0.005	0.005

The second assumption of IRT models is local independence which was investigated by calculating Q3 statistics. The findings were given in Table 2 below. Although there is no agreed threshold value to interpret Q3 values in the literature, Christensen and his colleagues (2017) stated that local independence is quite unlikely for values of 0.3 and lower. When 0.3 threshold value is taken when deciding whether or not item pairs show local dependence, the table's values implied a possible local dependence between items 1 and 3. This finding requires further investigation and needs to be cleared because local dependence was not observed for multiple pairs; this result was regarded not deteriorating the local independence assumption.

Table 2. Q3 Local Dependence Statistics between the Items of SWLS

Items #	item 1	item 2	item 3	item 4	item 5
item 1		-0.17	-0.32	-0.13	-0.16
item 2			-0.14	-0.32	-0.16
item 3				-0.28	-0.21
item 4					-0.12
item 5					

3.2. Fitting the GRM to SWLS

After checking the assumptions, the GRM was tested. According to the findings, it can be said that the fit of the unidimensional GRM model fits well [$M^2(5) = 8.86$, $p = 0.115$, RMSEA = 0.041 (95% CI = 0.029-0.083), SRMR = 0.029]. Also, reliability was again estimated in the context of IRT. The result showed that the IRT based empirical (marginal) reliability value of SWLS was 0.82.

Later, each item's fit to GRM was evaluated with Orlando and Thissen's (2003) S- χ^2 statistics. Significant S- χ^2 values are considered as a misfit for a given item. The values on item fit statistics and information amounts are given in Table 3 below. The results revealed that all the items' fit values were not significant at $p > 0.05$ level. Additionally, it was found that the highest amount of information was obtained from the 3rd item while the least amount of information obtained from the 5th item.

Table 3. S-X² Item Fit Statistics, Item Information Values and Total Information Value

Items	S- χ^2	df	p	Item Info.	Total Info.
item 1	62.43	52	0.153	6.54	31.34
item 2	43.57	57	0.905	5.93	
item 3	44.67	54	0.813	7.95	
item 4	71.46	56	0.080	6.56	
item 5	80.27	62	0.059	4.35	

The item parameters and IRT based factor loadings of the GRM were given in Table 4 below. As shown in the table, the factor loadings varied between 0.66 and 0.79 and the difficulty parameters varied between -3.07 and 2.44. The highest value of the difficulty values belonged to the 1st item and the lowest value belonged to the 4th item, the difficulty parameters of the items spread over a wide range of ability levels. Although the discrimination parameters are the only factor that determining the level of information an item provides, a wide range of difficulty parameter values is a prerequisite for these items to provide measurement accuracy across different levels of θ . The discrimination parameters of items varied between 1.51 and 2.19. According to Baker (2001), values between 1.35 and 1.69 have “high discriminating power” while values at and above 1.7 have “very high discriminating power”. Based on that criteria, it can be said that items 2 and 5 have a “high” level of discrimination while other items have a “very high” level of discrimination.

Table 4. Item Parameters and Factor Loadings Extracted From GRM

Item	F1	a1	d1	d2	d3	d4	d5	d6
item 1	0.73	1.80	-3.07	-2.07	-1.43	-0.59	0.18	2.06
item 2	0.70	1.67	-2.74	-1.48	-0.86	0.17	0.89	2.42
item 3	0.79	2.19	-2.19	-1.47	-0.94	-0.15	0.45	2.03
item 4	0.74	1.85	-2.79	-1.70	-0.98	-0.35	0.32	1.96
item 5	0.66	1.51	-1.54	-0.59	-0.13	0.64	1.34	2.44

In Figure 1 below, the category response functions (also known as item trace lines) of the items belonging to SWLS were given. These category functions allow the evaluation of each item's response options and show which categories are more likely to be preferred at which θ levels. The ideal response categories should be selected in a particular part of the ability being measured. Additionally, category response functions must not overlap and the order of them should not change. From this point of view, it shows that the probability of endorsement of all items for response category 3 and 5 never reached the highest probability at any point on the θ spectrum other response options even have the highest response probability in some range of abilities. This result implies that when the respondents select response categories of SWLS items, they probably find the adjacent options of the 3th and 5th response categories more attractive. These results put the existence of these categories into question.

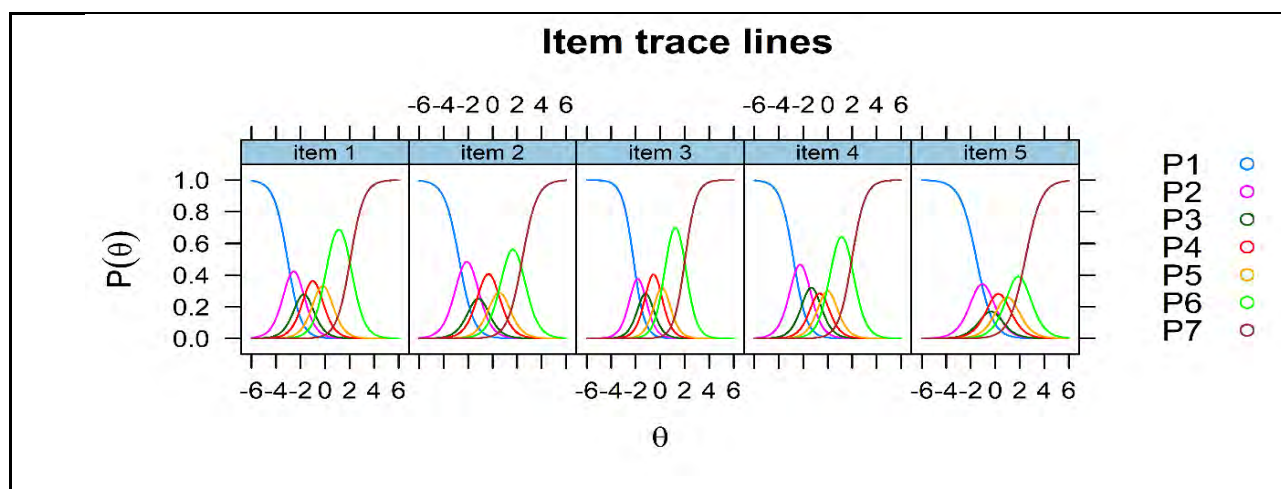


Figure 1. Category Response Functions and Information Traces of SWLS

Test information and reliability functions of SWLS were given in Figures 2a and 2b below. In the figure, the x-axis corresponds to the predicted life satisfaction score (θ), the left side of the y-axis shows the amount of information given by the test while the right side shows the amount of standard error. The information function's peak shows the level of ability that the SWLS yields the most information (enable the most precise estimation of ability). When the shape of the function was examined, the highest amount of information can be obtained in the range of -2 to 0 ability levels. On the other hand, the amount of information given in the range of -3 to 2 ability levels did not change considerably. This finding implies that the SWLS provides information in a wide range of ability level. As to the amount of standard errors, the figure showed that, standard errors minimised at the ability levels where the most information was yielded.

Another advantage of the IRT is that it can give the amount of reliability for different ability levels. The reliability function of the SWLS was shown in Figure 2a below. The reliability function and the information function goes fairly parallel. Figure 2b below also includes a reference line to provide information about the range of ability where reliability is greater than .8. The given reference line was drawn based on the acceptable level of reliability suggested by Shavelson (2004). Accordingly, it is seen that SWLS has a reliability of .8 and above in the range of -3 to 3 ability levels and the reliability level decreases for more extreme ability levels. Such a wide range of acceptable level of reliability can be regarded as a strength of SWLS.

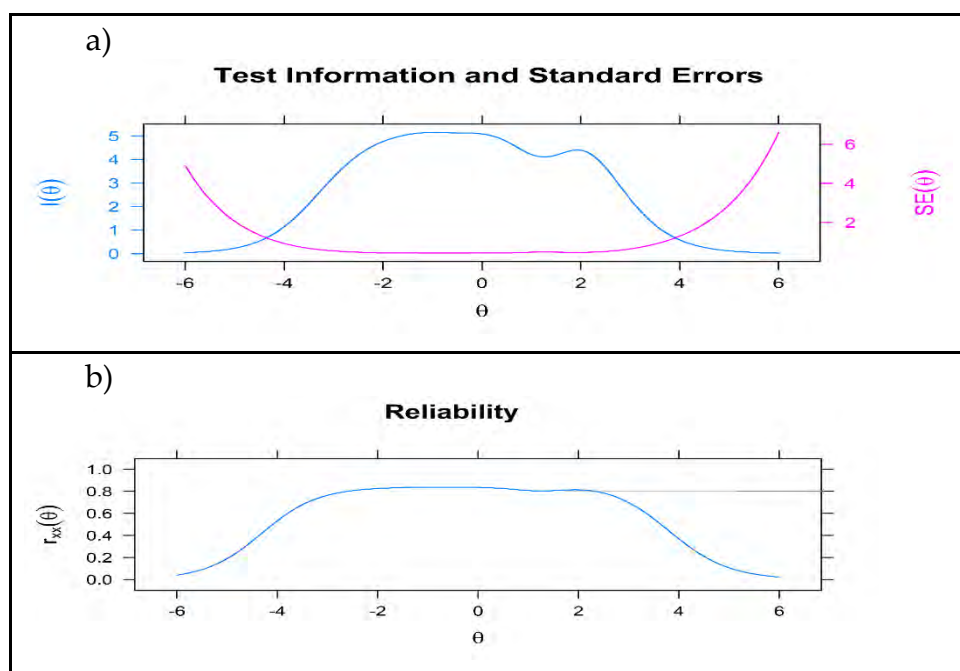


Figure 2. Test Information Function, Standard Error Function and Reliability Function of SWLS

4. Discussion

In this study, psychometrical properties of SWLS were assessed by IRT based approach with university students. The data were collected in a large Metropolitan. For preliminary analysis, both exploratory and confirmatory factor analysis were conducted to test the unidimensionality of SWLS and results support the unidimensional structure. Additionally, the obtained Q3 statistics provided satisfactory evidence of the local independence assumption. Furthermore, Lord's χ^2 statistics provide evidence that SLWS items did not show gender-based DIF. This finding is significant because it implies that the scale items function similarly for men and women and do not give biased estimations for different gender groups. This result also supports the validity of the SWLS.

Later, GRM was applied to the SWLS data and fit indices provided evidence that GRM can be used in future studies by the researchers. When the items' discrimination parameters were examined, it was seen that the items have "high" and "very high" discrimination values. Furthermore, it has been found that the predicted difficulty parameters of the items cover a wide range of ability spectrum. As a result, the SLWS reliability level did not fall below 0.8 in a wide range. In parallel, the standard error level remained low in the same range of ability. Although the reliability value obtained with the classical approach is 0.8, it is only an assumption that this level of reliability does not change across the ability spectrum. The results obtained with the IRT approach

provided evidence for this assumption; they showed that 0.8 or a higher level of reliability could be obtained for individuals for a wide range of ability. On the other hand, the results revealed that, for extremely higher and lower levels of ability, the reliability level of the SWLS drop.

Although CTT approach was used to assess validity and reliability of SWLS in different populations (i.e. Amtmann et al., 2017), no study benefited from using IRT approach to investigate psychometric properties of SWLS. In this respect, this study is the first one using IRT approach to evaluate the validity of SWLS. As cited in the first part, modern IRT techniques provide many advantages over CTT (Reeve & Fayers, 2005). In this way, we believe that the current study helps researchers obtain deeper insights into the item and measurement properties of SWLS. The findings presented in this study were similar to the findings obtained by Amtmann et al. (2017) in a way that, both studies provided evidence for the unidimensionality of SWLS.

Furthermore, in this study, it was found that the third and fifth response categories had a lower probability of endorsement compared to the adjacent response categories. The endorsement likelihoods of those categories are not higher than the alternative categories at any part of the talent trait spectrum. One of the possible causes of this disturbance is that participants may find it difficult to distinguish between adjacent category labels. In this case, it may be better to use category labels that can be more easily distinguished or reduce the number of response categories. Hence, it may be better to combine these categories with adjacent ones. Amtmann et al. (2017) also suggested that the seven response categories for SWLS are too high and should be reduced to five. Hence, future studies need to be carried out to see how the measurement precision will be affected when the response categories of the SWLS are reduced. Additionally, Amtmann recommended the removal of fifth item. This recommendation was partially supported in the current study as it provided the least amount of information, item fit statistics supported keeping of it in the scale. For this reason, instead of removing it from the scale, 5th item could be revised and the possible effect of this revision on the psychometric properties of SWLS could be studied.

5. Conclusion

It can be concluded that even the number of items of the SWLS is small; it can provide a high level of reliable measurement quality in a wide range of ability spectrum. On the other hand, it should be taken into consideration that these findings are limited because the sample size was not large, and the sample group consisted only of university students. In future studies, this study could be replicated with diverse sample groups.

In conclusion, readers should be aware of one important point that IRT approach's findings should not be perceived as an alternative to the CTT. Although the findings obtained by the IRT approach can be beneficial and help in gaining new insights, it is necessary to compare the findings obtained in CTT and IRT approaches. At the same time, beyond statistical findings, content experts' opinions regarding the practical impacts of any kind of revision made on SWLS should be considered. Readers should never forget that integrating the IRT into the classical approach would be more beneficial instead of replacing it. Finally, this study used the IRT approach, which included modern psychometric procedures to measure life satisfaction. Hoping that this study will contribute to the measurement and evaluation of life satisfaction considering the infrequent use of modern techniques in validating life satisfaction.

6. References

- Amtmann, D., Bocell, F. D., McMullen, K., Bamer, A. M., Johnson, K. L., Wiechman, S. A., & Schneider, J. C. (2020). Satisfaction with life over time in people with burn injury: a national institute on disability, independent living, and rehabilitation research burn model system study. *Archives of physical medicine and rehabilitation*, 101(1), S63-S70. <https://doi.org/10.1016/j.apmr.2017.09.119>
- Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied psychological measurement*, 2(4), 581-594. <https://doi.org/10.1177/014662167800200413>
- Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. <https://doi.org/10.1007/BF02293814>

- Appleton, S., & Song, L. (2008). Life satisfaction in urban China: components and determinants. *World Development*, 36(11), 2325-2340. <https://doi.org/10.1016/j.worlddev.2008.04.009>
- Baird, B. M., Lucas, R. E., & Donnellan, M. B. (2010). Life satisfaction across the lifespan: findings from two nationally representative panel studies. *Social indicators research*, 99(2), 183-203. <https://doi.org/10.1007/s11205-010-9584-9>
- Baker, F. B. (2001). *The basics of item response theory*. For full text: <http://ericae.net/irt/baker..>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M. and Novick, M.R., Eds., *Statistical theories of mental test scores*, Addison-Wesley, Reading, 397-479.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model* (2nd ed.). Lawrence Erlbaum.
- Meadow, H. L., Mentzer, J. T., Rahtz, D. R., & Sirgy, M. J. (1992). A life satisfaction measure based on judgment theory. *Social Indicators Research*, 26(1), 23-59. <https://doi.org/10.1007/BF00303824>
- Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for yen's q3: identification of local dependence in the rasch model using residual correlations. *Applied Psychological Measurement*, 41(3), 178-194. <https://doi.org/10.1177/0146621616677520>
- Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (Vol. 32). CRC press.
- Crane P.K., van Belle G., & Larson E.B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine*. 23(2), 241-256. <https://doi.org/10.1002/sim.1713>
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391-218. <https://doi.org/10.1177/0013164404266386>
- Dağlı, A., & Baysal, N. (2016). Yaşam doyumu ölçeğinin türkçe'ye uyarlanması: geçerlik ve güvenirlik çalışması. *Elektronik Sosyal Bilimler Dergisi*, 15(59), 1250-1262. <https://doi.org/10.17755/esosder.263229>
- DeMars, C. (2010). *Item response theory*. Oxford University Press.
- Diener, E. & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social Indicators Research*, 40(1), 189-216. <https://doi.org/10.1023/A:1006859511756>
- Diener, E., & Seligman, M. E. P. (2004). Beyond money: toward an economy of well-being. *Psychological Science in the Public Interest*, 5(1), 1-31. <https://doi.org/10.1111/j.0963-7214.2004.00501001.x>
- Diener, E., Emmons, R.A., Larsen, R.J. & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71-75. https://doi.org/10.1207/s15327752jpa4901_13
- Dorahy, M., J., Lewis, C., A., Schumaker, J., F., Akuamoah-Boateng, R., Duze, M., C. ve Sibiya, T., E. (2000). Depression and life satisfaction among Australian, Ghanaian, Nigerian, Northern Irish, and Swazi university students. *Journal of Social Behavior and Personality*, 15(4), 569-580.
- Durak, M., Senol-Durak, E. & Gencoz, T. (2010). Psychometric properties of the satisfaction with life scale among turkish university students, correctional officers, and elderly adults. *Social Indicators Research*. 99(3), 413-429. <https://doi.org/10.1007/s11205-010-9589-4>
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8(4), 341-349. <https://doi.org/10.1037/1040-3590.8.4.341>
- Funk, B. A. III, Huebner, E. S., & Valois, R. F. (2006). Reliability and validity of a brief life satisfaction scale with a high school sample. *Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being*, 7(1), 41-54. <https://doi.org/10.1007/s10902-005-0869-7>
- Hambleton, R. K., Shavelson, R. J., Webb, N. M., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
- Hattie J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139-164. <https://doi.org/10.1177/014662168500900204>

- Henard, D. H. (2000). Item response theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics*, (67-98). American Psychological Association publications.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum.
- Huebner, E.S. (1991). Initial development of the student's life satisfaction scale. *School Psychology International*, 12(3), 231-240. <https://doi.org/10.1177/0143034391123010>
- Iwasa, H., Kawaai, C., Gondo, Y., Inagaki, H., & Suzuki, T. (2006). Subjective well being as a predictor of all-cause mortality among middle-aged and elderly people living in an urban Japanese community: A sevenyear prospective cohort study. *Geriatrics & Gerontology International*, 6(4), 216-222. <https://doi.org/10.1111/j.1447-0594.2006.00351.x>
- Kang, T., & Chen, T. T. (2007). *An Investigation of the performance of the generalized s-x2 item-fit index for polytomous IRT models*. ACT Research Report Series, 2007-1. ACT, Inc.
- Köker, S. (1991). *Normal ve sorunlu ergenlerin yaşam doyumu düzeyinin karşılaştırılması*. Master Thesis, Ankara University, Institute of Social Sciences, Ankara.
- Köker, S. (1991). *Normal ve sorunlu ergenlerin yaşam doyumu düzeyinin karşılaştırılması* [Master Thesis, Ankara University]. Thesis Center of Higher Education.
- López-Ortega, M., Torres-Castro, S., & Rosas-Carrasco, O. (2016). Psychometric properties of the satisfaction with life scale (swls): secondary analysis of the Mexican health and aging study. *Health and Quality of Life Outcomes*, 14(1), 1-7. <https://doi.org/10.1186/s12955-016-0573-9>
- Lubin, B., & Van Whitlock, R. (2004). Psychometric properties of the brief life satisfaction scales. *Journal of clinical psychology*, 60(1), 11-27. <https://doi.org/10.1002/jclp.10190>
- Lucas-Carrasco, R., Den Oudsten, B. L., Eser, E., & Power, M. J. (2014). Using the satisfaction with life scale in people with parkinson's disease: a validation study in different european countries. *The Scientific World Journal*, <https://doi.org/10.1155/2014/680659>
- Margolis, S., Schwitzgebel, E., Ozer, D.J. and Lyubomirsky, S. (2018). A new measure of life satisfaction: the riverside life satisfaction scale. *Journal of Personality Assessment*. 101(6), 621-630. <https://doi.org/10.1080/00223891.2018.1464457>
- Masten, A. S. (2001). Ordinary magic: Resilience processes in development. *American Psychologist*, 56(3), 227-238.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174. <https://doi.org/10.1007/BF02296272>
- Matheny, K. B., Curlette, W. L., Aysan, F., Herrington, A., Gfroerer, C. A., Thompson, D., & Hamarat, E. (2002). Coping resources, perceived stress, and life satisfaction among Turkish and American university students. *International Journal of Stress Management*, 9(2), 81-97. <https://doi.org/10.1023/A:1014902719664>
- Maydeu-Olivares A, Drasgow F, Mead A.D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, 18(3), 245-256. <https://doi.org/10.1177/014662169401800305>
- Meadow H. L., J. T. Mentzer, D. R. Rahtz and M. J. Sirgy. (1992). A life satisfaction measure based on judgment theory, *Social Indicators Research*, 26(1), 23-59. <https://doi.org/10.1007/BF00303824>
- Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied psychological measurement*, 16(2), 159-176. <https://doi.org/10.1002/j.2333-8504.1992.tb01436.x>
- Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17(4), 351-363. <https://doi.org/10.1002/j.2333-8504.1993.tb01538.x>
- Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide*. Sixth Edition. Muthén & Muthén.
- Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691-692.

- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298. <https://doi.org/10.1177/0146621603027004004>
- Pavot, W., & Diener, E. (1993). Review of the Satisfaction with Life Scale. *Psychological Assessment, 5*(2), 164-172.
- Pavot, W., Diener, E., & Suh, E. (1998). The temporal satisfaction with life scale. *Journal of personality Assessment, 70*(2), 340-354. https://doi.org/10.1207/s15327752jpa7002_11
- R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software, 48*(6), 1-29.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*, Danish Institute for Educational Research.
- Sachs, J. (2003). Validation of the satisfaction with life scale in a sample of Hong Kong university students. *Psychologia, 46*(4), 225-234. <https://doi.org/10.2117/psysoc.2003.225>
- Samejima F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement, 34*(4), 139-139.
- Seung W. Choi, with contributions from Laura E. Gibbons and Paul K. Crane (2016). *lordif: Logistic Ordinal Regression Differential Item Functioning using IRT*. R package version 0.3-3. <https://CRAN.R-project.org/package=lordif>.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement, 27*(4), 361-370. <https://doi.org/10.1111/j.1745-3984.1990.tb00754.x>
- Vazquez, C., Duque, A., & Hervas, G. (2013). Satisfaction with life scale in a representative sample of Spanish adults: validation and normative data. *The Spanish journal of psychology, 16*(82), 1-15.
- Veenhoven, R. (1988). The Utility of Happiness. *Social Indicators Research, 20*(4), 333-354. <https://doi.org/10.1007/BF00302332>
- Zanon, C., Bardagi, M.P., Layous, K. et al. (2014). Validation of the satisfaction with life scale to Brazilians: evidences of measurement noninvariance across Brazil and US. *Social Indicators Research, 119*(1), 443-453. <https://doi.org/10.1007/s11205-013-0478-5>
- Zumbo, BD. (1999). *A handbook on the theory and methods of differential item functioning (dif): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.