International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 668–681

DOI: 10.21449/ijate.453337

Published at http://www.ijate.net http://dergipark.gov.tr/ijate Research Article

Gaining a Better Understanding of General Mattering Scale: An Application of Classical Test Theory and Item Response Theory

Halil Ibrahim Sari [1,*], **Mehmet Akif Karaman** [1]

[1] Kilis 7 Aralik University, Muallim Rifat Faculty of Education, Department of Educational Sciences, Kilis, Turkey

ARTICLE HISTORY

Received: 14 August 2018
Revised: 13 September 2018
Accepted: 24 September 2018

KEYWORDS

Classical test theory,
Item response theory,
Generalized partial model,
Item analysis,
General mattering scale

Abstract: The current study shows the applications of both classical test theory (CTT) and item response theory (IRT) to the psychology data. The study discusses item level analyses of General Mattering Scale produced by the two theories as well as strengths and weaknesses of both measurement approaches. The survey consisted of a total of five Likert-type items. Each student chose the best answers from the given categories (e.g., not at all, a little, somewhat, very much). We specifically run generalized partial credit IRT model. Overall, we discussed that while CTT provides comparatively superficial information, IRT allows gaining deeper insight into the test items and response categories. We concluded that the meaning of item properties differs in CTT and IRT, and this difference may lead to different interpretations. We aimed to encourage psychologists and counsellors to give more consideration to the IRT applications when assessing the psychometric features of the items.

1. INTRODUCTION

The purpose of many tests in the area of educational, psychological or behavioral measurement is to observe and explain human behaviors or feelings, and draw accurate inferences from given answers to test items. However, this is not straightforward task since latent variables are intended to measure unobservable or indirect constructs, such as feelings or emotions. Since the latent variable is measured by a set of items, how much one knows about the items in the test batteries determines how much he/she knows about latent variables. Consequently, how one analyzes, and then makes inferences about test items play a vital role. This study showed how inferences and interpretations made from the items intended to measure students' feelings about mattering would differ across different measurement theories.

There are two types of measurement theories used to describe the relationships between observed (e.g., items) and unobserved variables (e.g., attitude or latent ability). The first one is

CONTACT: Halil İbrahim SARI ✉ hisari87@gmail.com Kilis 7 Aralik University, Muallim Rifat Faculty of Education, Department of Educational Sciences, Kilis, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

the classical test theory (CTT). The CTT is a simple measurement theory which also known as true score theory. During the history of psychometrics dating back to early 1900s (Crocker & Algina, 1986), the CTT has been very popular, and intensively used by researchers and practitioners from a variety of fields including counseling and psychology (e.g. Cho, Drasgow, & Cao, 2015; Riemer & Kearns, 2010). This is because the CTT based calculations can be easily computed, even by hand. However, it has been extensively criticized due to producing population or sample dependent interpretations (Lord, 1980). This means that when the same survey is taken by two different groups, item properties such as discrimination and difficulty, hereby, inferences might differ across the groups. The second one is item response theory (IRT; Baker, 1992). IRT, also known as latent trait theory, is a strong psychometric theory. IRT explains the statistical relationships between test taker's ability level and the response given to that item by item response category characteristic curves. Unlike CTT, IRT allows to draw population independent interpretations as long as one has enough sample size (e.g., a minimum of 300). It also separately estimates the errors in the observed scores for all individuals.

Although IRT requires complicated calculations, due to the advantages over CTT, it is seen as the evolutionary measurement theory by practitioners and measurement theorists. Thus, it is very common to see its applications in many studies from a variety of areas including health education (Hays, Morales, & Reise, 2000), teacher education (Brodin, Fors, & Laksov, 2010), psychology (Lee, Krishnan, & Park, 2012), childhood education (Gomez & Vance, 2015), counseling (Riemer & Kearns, 2010) and more. However, unlike CTT, IRT based application examples are limited in the fields of counseling and psychology, and as compared to the CTT, a little consideration has been given to the IRT-based analyses by psychologists.

This study illustrated an application of both CTT and IRT approaches to the psychology related data, and compared item level analyses and interpretations produced by the two measurement theories. This collaboration aimed to empirically support the notion that IRT-based inferences would be more meaningful, hereby its applications should take more place in the field of psychology. Our goal was to draw researchers' attention to the IRT, and encourage counselors and psychologists to give more consideration to the IRT models in their research for a better understanding of their surveys.

1.1. Related Studies

Even though the CTT has been the dominantly used theory to assess the psychometric characteristics of the measurement tools, IRT-based applications has gained popularity in recent years. There are some good examples that used IRT approach when assessing the outcomes of the psychological scales. For example, Zanon, Hutz, Yoo and Hambleton (2016) used IRT-based calculations on affect scale that aimed to explore students' emotions and feelings as pleasant or unpleasant. They specifically used the graded response model, and discussed the highlights of IRT for psychological test development. Hamzeh and Fatima (2016) run graded response model, and discussed the findings of IRT-based applications when validating marital satisfaction scale. Riemer and Kearns (2010) compared both CTT and IRT-based findings and interpretations on youth counseling impact scale, and discussed the advantages of IRT over CTT. Another great example was provided by Turner, Betz, Edwards and Borgen (2017). They examined item quality of self-efficacy scale by using both CTT and IRT, and concluded that IRT-based calculations provided better insight to detect low and high quality items. Furthermore, Tasca et al., (2016) used both CTT and IRT when validating the shortened version of the therapeutic factors inventory form. Similarly, Locke et al., (2012) showed CTT and IRT based calculations, and discussed the strengths and weaknesses of the interpretations drawn across the two measurement theories when shortening the counseling center assessment of psychological symptoms scale. Lastly, Lee, Krishnan and Park (2012) run nominal response IRT model, which is a member of Rasch models. They analyzed psychometric properties of the children's depression inventory scale items, and reported item quality across the different age groups.

1.2. Theoretical Framework

When achieving the goal of measuring student feelings, typically Likert-type or polytomously scored items (aka multiple category items) are used. This is because polytomously scored items provide more information than dichotomously scored items (e.g., binary such as true/false, correct/incorrect, pass/fail; Embretson & Reise, 2000). Unlike dichotomously scored items (e.g., binary), polytomously scored items have more than two categories (e.g., strongly disagree, disagree, agree, and strongly agree). However, it is important to add that the choice of item format obviously depends on the purpose and consequences of the survey.

The CTT assumes that a test taker's test score is comprised of his/her "true" score and some measurement error. This is conceptualized as

$$X_{OBSERVEDTESTSCORE} = X_{TRUESCORE} + Error$$

The CTT assumes that the error in the observed score is constant across all test takers, meaning that everyone's true score is calculated by the same amount of error, which is unrealistic in practice. Another assumption is that there is no correlation between true scores and error (Crocker & Algina, 1986).

When items are dichotomously scored, item means represent item difficulty in CTT (i.e., proportion of examinees that answered an item correct). However, when the test items are polytomously scored, item means do not represent more than the average numerical mean score for an item. Item discrimination represents the ability to differentiate low proficiency and high proficiency examinees, and calculated by item-total correlations. If people get score low on an item also get low score on the test (or vice versa), this means that the item has better discrimination parameter.

In the IRT, there are two types of family modeling approaches. These are called cumulative (e.g., indirect) and adjacent (e.g., direct) models. Both type of models define step functions, the relationship between probability of reaching a certain category and latent variable, but differ in interpreting and calculating the steps. The cumulative models define step functions as the probability of scoring for cumulative score categories (e.g., probability of scoring more than 0, probability of scoring more than 1 etc.). Graded Response Model (GRM; Samejima, 1969) is a well-known example of polytomous IRT model that uses cumulative or indirect approach. The direct models, which is the interest in this study, define step functions as the probability of scoring for adjacent score categories (e.g., probability of scoring 1, for a person given that he/she already scored 0 or probability of scoring 2, given that person already scored 1). The Partial Credit Model (PCM; Masters, 1982) and the Generalized Partial Credit Model (GPCM; Muraki, 1992) are some examples of polytomous IRT models that use direct approach. If there are number of k response categories for an item, both PCM and GPCM specify m difficulty parameters (b), where m=k-1, unique to each step. However, while the GPCM specifies separate discrimination parameters for each item, and the PCM (i.e., also known as polytomous Rasch model) does not allow discrimination parameters to vary across the items, implying that the discrimination parameter equals to 1 for all items. It can be said that the two models are the nested models. The choice of the IRT model approach mainly depends on the theoretical and empirical considerations. One can refer to Huggins-Manley and Algina (2015) for more technical details about the polytomous item response theory.

In this study, we are particularly interested in direct or adjacent approach, and specifically run the GPCM due to empirical reasons (e.g., providing better model fit than PCM). One can refer to the Table 1 for the model-fit information for these two IRT models.

Table 1. *Item Response Theory Model-Fit Information*

Model	AIC	BIC	Log Likelihood	*df*	*p*
Partial Credit Model	18358.7	18436.6	-9164.3	15	
Generalized Partial Credit Model	18297.2	18405.0	-9128.6	20	<.001*

*There is a statistically significant difference between the nested models. Less restricted model (GPCM) fits better with the data.

The probability that an examinee scores the category *j* on item *i* is modeled by the GPCM as

$$P(Y=j|\theta)=\frac{exp^{a_i(\theta-b_{ij})}}{\sum_{k=0}^{m}\sum_{j=0}^{k}exp^{a_i(\theta-b_{ij})}}$$

Where *k* is the number of response options, *m* is the number of the steps, and *m*= *k*-1, a_i is the difficulty parameter and b_i is the difficulty parameter for item *i* on the *j*th step. The difficulty of a step (aka step parameter) represents the point at which a person has an equal chance or probability of scoring *j* or *j*+1 (advancement on the *j*th step). It is important to note that step parameters (e.g., b_1, b_2, b_3) are the interaction points of the two adjacent category response curves. The item response category characteristic curve given in Figure 1 visually shows an example of an item with four response options where b_1=-2.28, b_2=-0.96, b_3=0.85. The discrimination parameter determines the steepness of the category characteristic curves. As the discrimination of an item increases, the category response curves become steeper; hereby probability of selecting a response option rapidly changes.

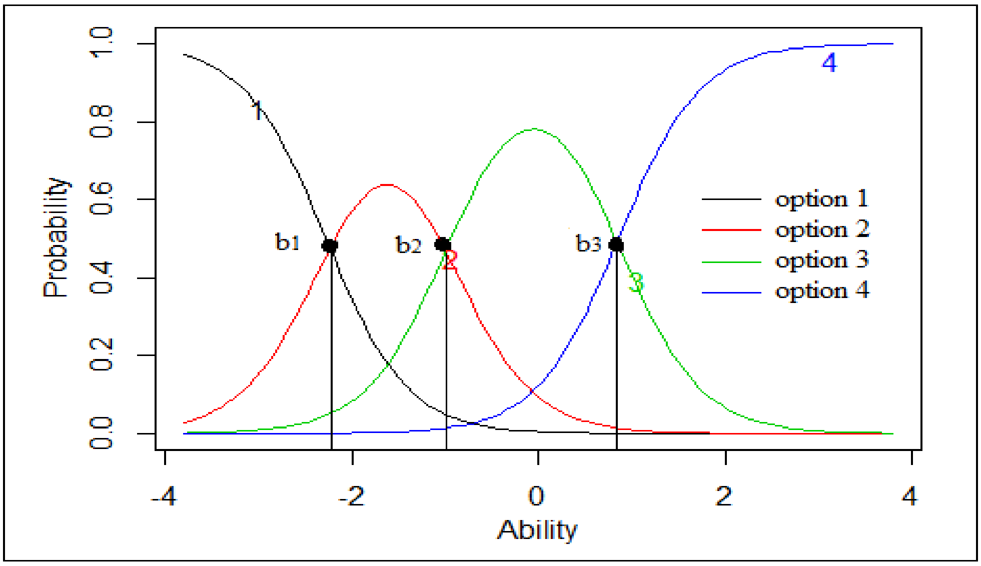

Figure 1. An example of item response category characteristic curve for a polytomously scored item.

1.3. Mattering

Many researchers in the fields of educational, social, and counseling psychology conducted studies on the concept of mattering and its relationship with other variables (Cha, 2016; Connolly & Myers, 2003; Flett, Galfi-Pechenkov, Molnar, Hewitt, & Goldstein, 2012; Rayle, 2006; Tovar, Simon, & Lee, 2009). However, no recent studies have explored this construct with Turkish participants or in the Turkish society. Rosenberg and McCullough conceptualized the term of mattering as a sense of belonging in 1980s (Tovar, 2013). After reviewing many research databases (e.g. PsychInfo, Academic Search Complete, Google Scholar, ULAKBIM) we have found that first time mattering was studied by Demir, Özen, and Doğan in 2012 with a group of Turkish participants. In other words, the term took attention of Turkish researchers 32 years after it was introduced by Rosenberg and McCullough. There are a few reasons that researchers in Turkey did not show sufficient interest in mattering.

First, Rosenberg and McCullough (1981) described mattering as one’s sense of belonging to his/her immediate surroundings and society. We have found that there are many studies conducted on sense of belonging in the Turkish literature. This could be one of the reasons that researchers focused on the core concept of mattering but neglected the big picture- mattering. The second reason could be the lack of mattering measurements available in Turkish. The study conducted by Demir et al. (2012) used Mattering to Others Questionnaire (Marshall, 2001); however, the instrument is not available in Turkish and we assume that it was translated by Demir and colleagues for the study they conducted in 2012. One problem with this is the lack of information of the psychometric properties of the instrument they used because there is no adequate information about the validation process in their study. At our best knowledge, Haktanır, Lenz, Can, and Watson (2016) developed, and validated the first mattering measure which was General Mattering Scale (Marcus, 1991) into Turkish.

In parallel to the current study’s aim, we believe that mattering will take more attention and both researchers and practitioners will benefit from its potential value. Moreover, measurements of mattering in different languages will help to conceptualize “our subjective perception and interpretation that we make a difference to others in our lives” (Tovar, 2013, p. 42). As stated by Rosenberg and McCullough (1981), three elements of mattering, which are attention, importance, and dependence, have continued to constitute the theoretical base of measurements and explain external validation of a person by others. This external validation comes true at the interpersonal and societal levels (Rosenberg & McCullough, 1981). The interpersonal dimension indicates individuals’ level of mattering to people in their lives. In other words, the feeling of mattering they get from close relationships, such as, parents, friends, teachers. On the other hand, the societal dimension includes individuals’ perception of mattering toward outer world, such as, schools, governmental institutions, religious institutions. As the number of mattering studies which have been conducted in different cultures increase, our understanding level of the concept and related variables increase.

2. METHOD

2.1. Participants and Data Collection

The relevant university ethics board approved the study, the data were collected from volunteer students attending a state university in southeast region of Turkey during 2017-2018 academic year. There were 1644 undergraduate students from five different faculties. Twenty-one cases were removed from the data file since participants left the instrument items blank. This ended the data with 1623 participants. Of the participants, 59.5% were female (n= 966) and 39.5% were male (n= 643), 14 participants failed to respond this demographic query. The ages of respondents were between 17 and 39, the average age of the total participants was 21.4.

Participants reported their academic levels as freshmen (n = 372, 23%), sophomores (n = 439, 27%), juniors (n = 524, 32.3%), and seniors (n=279, 17.2%). Nine participants failed to respond to this demographic query.

2.2. Measure

In the current study, we used the General Mattering Scale- Turkish Version (GMS-TR; Haktanir et al., 2016) to collect data. The GMS (Marcus, 1991) was developed to assess the degree individuals believe how they are important to others. This 5-point Likert-type assessment yields a single scale score based on participant responses that range from Very Much to Not at All. Possible scores on the scale range from five to 20, with higher scores being indicative of a greater perception of mattering. Mattering is accounted for by participant responses to items such as "How important do you feel you are to other people?" and "How interested are people generally in what you have to say?" Haktanır et al. (2016) reported moderate Cronbach's alpha coefficients of .74. For the original version, Rayle and Myers (2004) reported alpha coefficients between .74 and .86 among college students. For the current study, we calculated a Cronbach's alpha of .76. The survey is given in the Appendix.

2.3. Analysis Procedure

Since the IRT models presented above assume dimensionality (e.g., measuring one and only one construct), we first checked the dimensionality of the data to ensure there was a single construct of interest, named general mattering. Specifically, we run a confirmatory factor analysis (CFA) with five items in AMOS (Arbuckle, 2014). The fitted model that we run was shown in Figure 2. For the CTT interpretations, we calculated the means for each response category across all items, CTT based-observed score distribution, and item level means and discrimination parameters. We also calculated CTT-based reliability index known as Cronbach's alpha. For the IRT interpretations, we run the GPCM, and calculated similar parameters. For the reliability measures in the IRT context, we calculated item and test information function which corresponds to the reliability index in IRT. We used the SPSS version 22.0 (IBM, 2013) for the CTT calculations, and the "ltm" package (Rizopoulos, 2017) in R Software (R Core Team, 2016) for the IRT calculations.

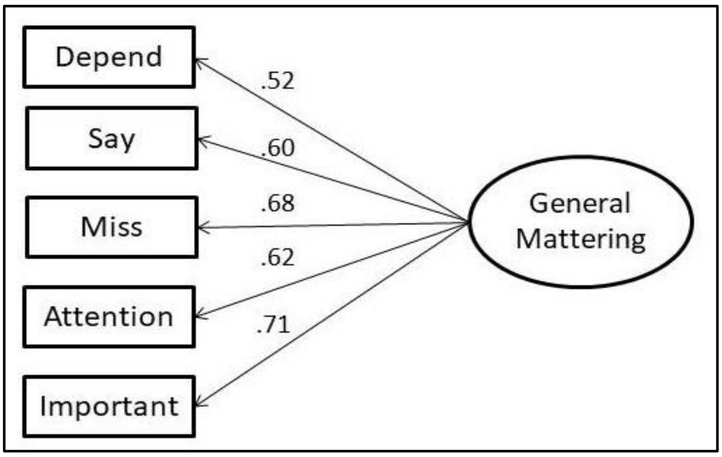

Figure 2. The specified CFA model of general mattering

3. RESULTS

The results of CFA indicated that the single factor model of general mattering scale (see Figure 2) had a strong fit; $\chi^2(5)= 47.22$, $p< .05$; GFI= .99, CFI= .97, TLI= .95, RMSEA= .07, and SRMR= .03. This shows that we met the unidimensionality assumption. Item and scale level findings produced by the CTT and IRT are discussed in the following paragraphs.

3.1. CTT Based Findings

Item means, standard deviations, item discrimination parameters and Cronbach's alpha statistics if item deleted across the five items, and scale level statistics produced under the CTT are given in Table 2. The mean for the item related to depend were the highest. It is obvious that there were a lot of students picked "somewhat" or "very much" on this item. The lowest mean was found for the item related to the importance. Item means for the remaining four items ranged from 2.66 to 2.91. This implies that on these four items, students generally picked the two middle options- "a little" and "somewhat". Based on the scale level statistics given at the bottom of Table 2, on average, a student got a score of 2.88 out of 4 for an item. Hence, it is safe to say that students mostly believe that the degree to which others matter to them was close to somewhat. The descriptive statistics for the response categories given in Table 3 also support this. The discrimination parameter, the power of differentiating the respondents, was the lowest and highest for the items related to the depend, and important, respectively. This means that the depend and important items were the most and least effective items to distinguish the students having low and high general mattering score in total, respectively. The discriminations for all items in the scale were higher than acceptable rates (e.g., 0.4). The scale level (e.g., across the five items) Cronbach's alpha measure was 0.76 which is not very high but at acceptable rate. It was also found that deleting any item would increase the alpha value. However, it was seen while the effect of deleting the first four items on the Cronbach's alpha would be more serious and, the effect of removing the depend item from analysis would be negligible.

Table 2. *Item and scale level statistics produced under classical test theory model*

Item Level	Item Mean	Standard Deviation	Item-Total Correlation (Discrimination)	Cronbach's Alpha If item deleted
Important	2.66	0.92	.59	.70
Attention	2.75	0.87	.52	.72
Miss	2.88	0.96	.57	.71
Say	2.91	0.87	.53	.72
Depend	3.18	0.87	.45	.75
Scale Level	Mean	Standard Deviation	Variance	Cronbach's Alpha
	2.88	0.64	0.41	0.76

Table 3. *Descriptive statistics for the response options*

	Response Option			
Item related to the	Not At All	A Little	Somewhat	Very Much
Important	235 (14.6%)	344 (21.4%)	751 (46.8%)	276 (17.2%)
Attention	140 (8.7%)	433 (27.0%)	706 (44.0%)	325 (20.3%)
Miss	164 (10.3%)	350 (22.1%)	581 (36.7%)	490 (30.9%)
Say	120 (7.5%)	341 (21.2%)	706 (43.9%)	441 (27.4%)
Depend	102 (6.3%)	207 (12.8%)	594 (36.8%)	711 (44.1%)

Note: Percentages are the valid percentages after removing missing cases

3.2. IRT Based Findings

The item parameters for the five items under the generalized partial credit model are given in Table 4, and the item response category characteristic curves that show the probability of selecting each response option are shown in Figure 3. By looking at the Table 4, we saw that b_1 parameter (e.g., first step parameter) was the highest for the important item, meaning that the chance of feeling "a little important to other people" for a person who scored "not at all" was more difficult compared to items. This also means that someone with θ= -1.06 has an equal chance of picking the response category 1 (Not at all) and response category 2 (A little) on this item. In other words, any student with ability score of less than θ = -1.06 will more likely pick "not at all", higher than θ =-1.06 will more likely pick "a little". Whereas, it was the lowest for the say item, meaning that someone with θ = -1.86 has an equal chance of picking "not at all" and "a little". As told before, this is because the corresponding value of the ability scale at the interaction point of the curves for these two options was -1.86 for the say item. Thus, we can conclude that stepping from the response option "not at all" to "a little" is easier or requires less amount of general mattering ability score on the say item than any other items. The b_1 parameters for the remaining items can be interpreted similarly.

Table 4. *Item parameters produced by the Generalized Partial Credit Model*

	Step Parameter			Discrimination
Item related to the	b_1	b_2	b_3	a
Important	-1.06	-0.64	1.24	1.64
Attention	-1.81	-0.60	1.17	1.16
Miss	-1.46	-0.65	0.53	1.37
Say	-1.86	-0.94	0.81	1.10
Depend	-1.79	-1.66	-0.05	0.83

When we looked at the transitions from the option "a little" to the option "somewhat", we saw that b_2 parameter was the lowest for the depend item, meaning that someone with θ =-1.66 has an equal chance of picking both middle categories on this item. When the latent ability score is higher than θ =-1.66, the likelihood of picking "somewhat" increases. Whereas, it was the highest for the attention item, meaning that someone with a theta of -0.60 has an equal chance of picking the both categories. More specifically, any person with ability score of less than θ = -0.60 (but less than θ =-1.81 because b_1 for this item is –1.81) would more likely pick the response option of a little. To sum up, the chance of stepping from a little to somewhat requires less and higher amount of general mattering ability score on the depend and attention items, respectively.

We found that b_3 parameter was the lowest for the depend item, and highest for the important item. This means that the transitions from "somewhat" to "very much" for students who already selected "somewhat" requires lowest and highest general mattering latent trait score for the depend and important items, respectively.

We also noticed that the probability of selecting "a little" was the lowest on any point of the ability scale for the depend item (e.g., the curve numbered 2 was always under the other curves on any points of the scale). This means that regardless of the amount of general ability score, students always tented to select other three options on the depend item. This was also true for the other items. This is because the likelihood of selecting the option "a little" was higher just for a limited range of ability scores (i.e., the curve numbered 2 was above the other curves across a short range on the ability scale). Thus, we can conclude that the option "a little" was the least efficient option, especially for the depend item, and regardless of the magnitude of the general mattering ability score, the students always tended select the other options.

In terms of item discriminations, the best discriminating item was important item, and the worst discriminating item was depend item. When we look at the Figure 3, it is clear that item category curves were steeper on the important item, and flatter on the depend item.

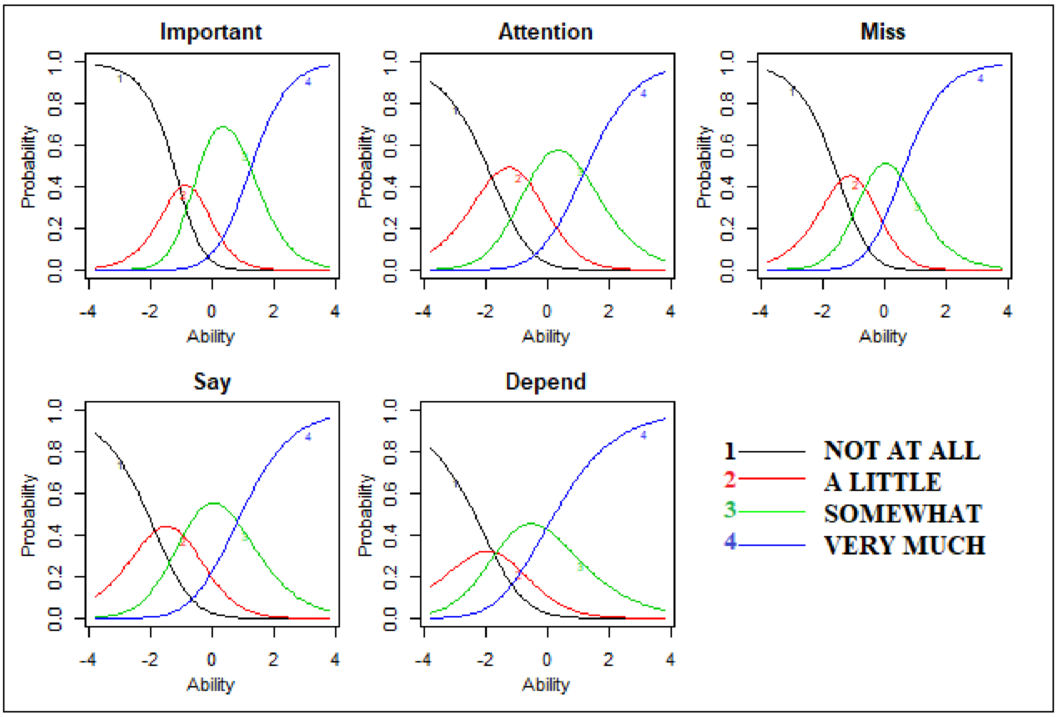

Figure 3. Item response category characteristic curves for all items.

We also provided item and test or scale level information function that shows (1) the amount of information an item or a scale provides across the different points on the ability scale, and (2) the point where the information peaks, and showed in Figure 4. Based on the Figure 4, the first finding was that item information function always peaks around the theta point of -1 for all items. This means that the test items were more appropriate for the students having ability score of -1. In other words, the latent ability scores were more precisely or accurately calculated for the students with the ability level of around -1. The test level information chart in the same figure also supports this. The second finding was that the depend item provides the lowest information, and important item provides the highest amount of information for the students across the ability levels. This means that the depend and important items were the worst and best items, respectively, to be able to measure general mattering ability with a less amount of measurement error.

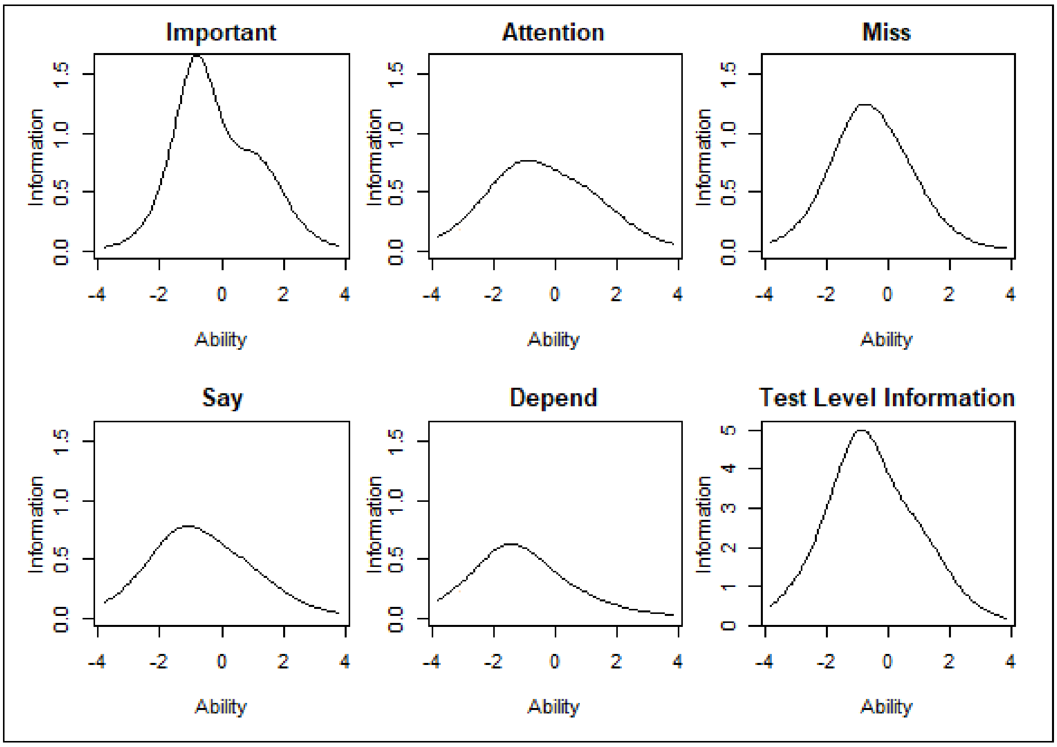

Figure 4. Total item information functions for the five items.

4. DISCUSSION

More likely due to computational easiness, classical test theory is more commonly used by the psychologists and counselors. However, because of producing sample dependent outcomes, it does not always allow researchers and practitioners to generalize findings to other facets. Furthermore, due to its test-focused feature (e.g., focusing total score), CTT does not tell much about item characteristics in detail and how persons are well on specific items. Whereas, item response theory can easily overcome these problems and can provide more meaningful and concrete interpretations (de Ayala, 2009). However, there is a lack of item response theory applications in the area of psychological measurement. This study showed an application of both classical test theory and item response theory applications to the counseling data, and presented the inferences made by the two.

Based on the CTT findings, for all items, students generally selected the second and third response options (e.g., a little and somewhat), but always the lowest percentage of the students picked the first response option (e.g., not at all). The distribution of selected response options was relatively closer on the important item which made it the most discriminating item, but varied on the depend item which led it to have the lowest discrimination parameter. Thus, it is safe to conclude that the in terms of the contribution to the scale for a better measurement, the depend item was the least effective item.

Item response theory also found that that the depend item was the most problematic item in the scale. This is because, this item provided the lowest amount of item information across the different ability levels, and had the lowest discrimination power. Moreover, the chance of reaching the highest level of response option (e.g., very much) was much easier on this item which was inconsistent with the other items. We believe that it is interesting to see that students at this age easily believe that people depend on them very much. There could be two possible explanations of this finding. First, from a psychological perspective, this means that it was more significant to be considered by others than someone else being depend on them. This was consistent with previous research conducted by Karaman, Balkin, and Juhnke (2018). Researchers conducted a life balance-related study with 453 Turkish participants, and indicated that there was a positive relationship between feeling important by others and a balanced life. In other words, participants concerned whether others cared about them or not. Second explanation of believing people depend on them very much could be that this was more likely because students in the current study did not have a better understanding of this item, and had difficulty interpreting. In order to eliminate the possible confusion, it would be beneficial to play with the structure of the item or to provide additional explanation in a parenthesis. On the other hand, the best item was the important item due to providing the highest amount of information and having the highest discrimination power. Also, the scale was more effective to measure students having general ability score of around -1. However, it was not easy to make these inferences by the CTT.

Furthermore, the IRT detected that the option “a little” is the least effective option, especially for the depend item. This was due to the fact that, the item category response curve for a little response option always stayed under the other response curves. Hence, the students always tended to select the other options, regardless of the magnitude of latent trait score. This might be because of that the students did not distinguish between the response options “a little” and “somewhat” when answering the items. We have concluded that this response option should be revised or removed.

This study does not argue against CTT-based calculations. Rather, this work simply shows how the meaning of item parameters would differ in two measurement approaches and how these differences may yield different interpretations. We aim to encourage psychologists and counselors to give more consideration to the item response theory applications when assessing the psychometric features of the items in the scales in the educational and counseling areas, and draw their attention to the IRT. This is because; as shown in the paper, IRT would provide more information about the items and response options. As the final note, the choice of measurement approach could be particularly very important when validating a psychometric tool.

In this work, we run GPCM only, but other studies should also examine properties of the items under other polytomously scored item response theory models using direct approach such as direct PCM or GRM.

ORCID

Halil İbrahim SARI http://orcid.org/0000-0001-7506-9000

5. REFERENCES

Arbuckle, J. L. (2014). *Amos 23.0 user's guide.* Chicago, IL: IBM SPSS.

Baker F. (1992). *Item response theory*. New York, NY: Marcel Dekker, INC.

Brodin, U., Fors, U., & Laksov, K. B. (2010). The application of item response theory on a teaching strategy profile questionnaire. *BMC medical education, 10,* 14-14-doi:10.1186/1472-6920-10-14

Cha, M. (2016). The mediation effect of mattering and self-esteem in the relationship between socially prescribed perfectionism and depression: Based on the social disconnection model. *Personality and Individual Differences, 88,* 148-159. doi:10.1016/j.paid.2015.09.008

Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological Assessment, 27,* 1241–1252. doi:10.1037/pas0000132

Connolly, K. M., & Myers, J. E. (2003). Wellness and mattering: The role of holistic factors in job satisfaction. *Journal of Employment Counseling, 40*, 152-160. doi:10.1002/j.2161-1920.2003.tb00866.x

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston.

de Ayala, R. J. (2009*). The theory and practice of item response theory*. New York, NY: Guilford.

Demir, M., Özen, A. & Doğan, A. (2012). Friendship, perceived mattering and happiness: A study of American and Turkish college students. *The Journal of Social Psychology, 152,* 659-664. doi: 10.1080/00224545.2011.650237

Dodeen, H., & Al-Darmaki, F. (2016). The application of item response theory in developing and validating a shortened version of the Emirate Marital Satisfaction Scale. *Psychological Assessment, 28,* 1625-1633. doi:10.1037/pas0000296.supp

Embretson S.E, & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, IN: Lawrence Erlbaum Associates, Inc.

Flett, G. L., Galfi-Pechenkov, I., Molnar, D. S., Hewitt, P. L., & Goldstein, A. L. (2012). Perfectionism, mattering, and depression: A mediational analysis. *Personality and Individual Differences*, *52*, 828-832. doi:10.1016/j.paid.2011.12.041

Gomez, R., & Vance, A. (2015). Item response theory properties of the internalizing disorders in adolescents. *Journal of Childhood & Developmental Disorders, 1,* 1-10. doi:10.4172/2472-1786.100006

Haktanir, A., Lenz, A. S., Can, N., & Watson, J. C. (2016). Development and evaluation of Turkish language versions of three positive psychology assessments. *International Journal for the Advancement of Counselling, 38,* 286-297. doi:10.1007/s10447-016-9272-9

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38,* II28-II42. doi:10.1097/00005650-200009002-00007

Huggins-Manley, A. C., & Algina, J. (2015). The partial credit model and generalized partial credit model as constrained nominal response models, with applications in M Plus. *Structural Equation Modeling: A Multidisciplinary Journal, 22,* 308-318. doi:10.1080/10705511.2014.937374

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp.

Karaman, M. A., Balkin, R. S., & Juhnke, G. A. (2018). Turkish adaptation of the Juhnke–Balkin Life Balance Inventory. *Measurement and Evaluation in Counseling and Development, 51*, 141–150. doi:10.1080/07481756.2017.1308226

Lee, Y. S., Krishnan, A., & Park, Y. S. (2012). Psychometric properties of the Children's Depression Inventory: An item response theory analysis across age in a nonclinical, longitudinal, adolescent sample. *Measurement and Evaluation in Counseling and Development*, *45*, 84-100. doi: 10.1177/0748175611428329

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P. W., Hayes, J. A., Castonguay, L. G., ... & Lin, Y. C. (2012). Development and initial validation of the Counseling Center Assessment of Psychological Symptoms–34. *Measurement and Evaluation in Counseling and Development*, *45*, 151-169. doi: 10.1177/0748175611432642

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum.

Marcus, F. M. (1991). *Mattering: Its measurement and theoretical significance for social psychology*. Paper presented at the annual meeting of the Eastern Sociological Association, Cincinnati.

Masters, G. (1982). A Rasch model for partial credit scoring. Psychometrika, *47*, 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Rayle, A. D. (2006). Mattering to others: Implications for the counseling relationship. *Journal of Counseling & Development*, *84*, 483-487. doi:10.1002/j.1556-6678.2006.tb00432.x

R Core Team (2016). R: A language and environment for statistical computing [computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Riemer, M., & Kearns, M. A. (2010). Description and psychometric evaluation of the Youth Counseling Impact Scale. *Psychological Assessment*, *22*, 259-268. doi:10.1037/a0018507

Rizopoulos, D. (2017). Package "ltm", Latent Trait Models under IRT. Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Rosenberg, M., & McCullough, B. C. (1981). Mattering: Inferred significance and mental health among adolescents. *Research in Community & Mental Health*, *2*, 163-182.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*, 1–97. doi:10.1007/bf03372160

Tasca, G. A., Cabrera, C., Kristjansson, E., MacNair-Semands, R., Joyce, A. S., & Ogrodniczuk, J. S. (2016). The therapeutic factor inventory-8: Using item response theory to create a brief scale for continuous process monitoring for group psychotherapy. *Psychotherapy Research*, *26*, 131-145. doi:10.1080/10503307.2014.963729

Tovar, E., Simon, M. A., & Lee, H. B. (2009). Development and validation of the college mattering inventory with diverse urban college students. *Measurement and Evaluation in Counseling and Development*, *42*, 154-178. doi:10.1177/0748175609344091

Tovar, E. (2013). *A conceptual model on the impact of mattering, sense of belonging, engagement/involvement, and socio-academic integrative experiences on community college students' intent to persist* (Doctoral Dissertation). Retrieved from CGU Theses & Dissertations (Paper 81). doi: 10.5642/cguetd/81

Turner, B. M., Betz, N. E., Edwards, M. C., & Borgen, F. H. (2010). Psychometric examination of an inventory of self-efficacy for the Holland vocational themes using item response theory. *Measurement and Evaluation in Counseling and Development*, *43*, 188-198. doi:10.1177/0748175610384810

Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*, 1-10. doi:10.1186/s41155-016-0040-x

APPENDIX

The following five questions are designed to measure the degree to which you believe that you matter to others. Please circle the appropriate response for what YOU believe.

GENERAL MATTERING SCALE	**NOT AT ALL**	**A LITTLE**	**SOMEWHAT**	**VERY MUCH**
1. How *important* do you feel you are to other people?	**1**	**2**	**3**	**4**
2. How much do you feel other people pay *attention* to you?	**1**	**2**	**3**	**4**
3. How much do you feel others would *miss* you if you went away?	**1**	**2**	**3**	**4**
4. How interested are people generally in what you have to *say*?	**1**	**2**	**3**	**4**
5. How much do people *depend* on you?	**1**	**2**	**3**	**4**