Examining Validity of the Candidate Notification Scale for Gifted Children by Mokken Scale Analysis

Özge Bıkmaz Bilgen[1,*]

[1]Department of Educational Measurement and Evaluation, Adnan Menderes University, Aydin, Turkey

*Correspondence: Department of Educational Measurement and Evaluation, Adnan Menderes University, Aydin, Turkey. E-mail: ozgebikmaz86@hotmail.com

Received: October 17, 2020 Accepted: November 24, 2020 Online Published: December 16, 2020

doi:10.5430/wje.v10n6p44 URL: https://doi.org/10.5430/wje.v10n6p44

Abstract

The purpose of this study is to examine the validity of the scale for identifying gifted children, whose validity was proven by exploratory, confirmatory factor analysis, and whose reliability was proven the Cronbach alpha coefficient for identifying children in the 3-6 age group, using Mokken scaling based on nonparametric item response theory. The study group of the research consists of 253 students. As a result of the analysis of the 13-item 3-dimensional scale (*above average ability*, *creativity* and *task commitment*) with Monotone homogeneity Model, it is seen that a one-dimensional structure is obtained and the model fits when analyzed as a three-dimensional construct. For model data fit, when the discrimination and reliability values of the items are examined, it can be said that the one-dimensional structure of the scale is at an acceptable level, and relatively higher values are obtained for each criterion in the three-dimensional structure compared to the one-dimensional structure. Based on the findings, it can be interpreted that a parallel result was obtained in the validity study based on non-parametric item response theory of the scale, which was developed based on confirmatory factor analysis.

Keywords: gifted children, mokken scale analysis, monotone homogeneity model

1. Introduction

Nowadays, some ways are followed in identifying gifted children. The most important element of reaching the right children through the paths followed is the measurement tools used to identify children as gifted. The validity and reliability of scores obtained with the measurement tool, which is the condition for measurement every feature correctly, is also an important point in identifying giftedness.

The framework of giftedness has been drawn by different researchers. Regarding it as a multi-dimensional construct, Renzulli (2000) explains giftedness with a triad construct as general and special ability level, task commitment and creativity. According to Renzulli' three-ring conception of giftedness these three features are to intersect in order for an individual to be considered as gifted (Renzulli, 2000). In order for a child to be considered gifted, his / her intelligence, motivation and creativity are expected to be high. Based on the theory, behaviors such as asking questions, coming up with new ideas or completing tasks can be considered as indicators of giftedness.

Identifying gifted children at an early age allows their education to be in line with their abilities. The method of identify them at an early age can be made based on observation. Teachers play an effective role in the realization of this observation for families and children attending pre-school institutions (Bildiren, 2018). Making the observation free from personal judgments, based on valid and reliable measurement tools is necessary for correct identification. In this context, the need to examine the measurement tools used for this purpose comes to the fore.

Identifying gifted children has become increasingly important in recent years. The need for practical measurement tools to be used in the candidacy stage (Preiffer, 2015), which is the first step in identifying children as gifted, is on the agenda. Classical Test Theory, the origin of which is based on the studies of Spearman (1905), is used in developing measurement tools and evaluating the measurement results (Crocker & Algina, 1986). Despite its widespread use, it is stated that the classical test theory has some shortcomings. That the item and test statistics are dependent on the sample from which they are obtained, individuals' abilities are dependent on the items administered to them, the reliability of the test depends on the sample obtained etc. can be listed among these shortcomings

(Crocker & Algina, 1986; Hambleton, Swaminathan, & Rogers, 1991; De Ayala, 2009; Hambleton & Jones, 1993). As an alternative to the shortcomings of the Classical Test Theory, the Item Response Theory (IRT) was developed at the end of the 1930s (De Ayala, 2009). And there are many possible construct domains to which IRT may be applied.

In the Item Response Theory, it is thought that the performance of individuals in responding to the items in the test is the ability or trait to be measured with that test. The relationship between individuals' abilities (unobservable) and their responses to the items (observable) described by a mathematical function called item response functions (Hambleton, Swaminathan, 1985). The shape of the item characteristic curves obtained based on the item response function reveals the relationship between the change in ability level and the probability of answering correctly. It is seen as an advantage that IRT is not dependent on samples and items in its estimates (Embretson & Reise, 2000). In this context, it can be said that it is an effective theory that is based on ability identification and can be used in the identification of gifted individuals.

In addition to the advantages of item response theory, large sampling and a large item pool are required in order for the results to be obtained with the theory to be qualified (Demars, 2010). Besides, item response theory has assumptions such as unidimensionality, local independence, and normality (Hambleton & Swaminathan, 1985). It is difficult for researchers to meet the relevant assumptions and to provide the conditions in order to make estimates with IRT. The importance of conducting studies with a large sample and a large item pool is undeniable, but it can be said that gifted children have a small sample among all children. When the aim is to work with a group such as gifted students that shows accumulation in a high score in a certain ability or characteristic, it is seen that the scores obtained from the individuals diverge from the normal distribution, in other words, they become skewed. When the assumptions are met and it is desired to work with IRT with which qualified results are obtained under certain conditions, but when the data is far from the normal distribution or is obtained from a small sample, techniques that can be used in the proof of validity are needed.

While performing validity and reliability studies with IRT, conditions such as having a certain sample size of the obtained data and fulfilling normal distribution conditions should be provided in order for the results to be less erroneous (De Ayala, 2009). The requirement to meet the assumptions and meet the conditions for IRT led researchers to develop non-parametric IRT models that provide ease of application in short tests and small samples (Junker & Sijtsma, 2001). It is stated that non-parametric models are more flexible in terms of assumptions and conditions than parametric models (Sijtsma and Junker, 2006). The important advantage of non (parametric) IRT models is that they relax some of the strict assumptions of parametric models (Sijitsma & Molenaar, 2002). Mokken models (monotone homogeneity and double monotonicity model) are examined under non (parametric) IRT (NIRT). These models ease the certain shapes (like lojistic et.) of item characteristic curve, which shows the relationship between latent trait and response. In the context of polytomous items the IRF replaced by the item step characteristic function (ISRF). ISRF explains the relationship between probability that the item step score and the latent trait.

Although NIRT models ease some limitations of parametric models with the shape of the ISRF, there are still some assumptions that should be check in Monotone Homogeneity Model (Sijitsma & Molenaar, 2002). Unidimensionality and local independence are the basic assumptions in estimation with nonparametric IRT as in parametric models. In addition, monotonicity and non-intersecting item step response functions can be considered. The assumption of unidimensionality means that all of the items measure one construct. The second important assumption is local independence which means that the response of the individual taking the test is not influenced by his/her other responses to items in the same test. The third assumption is monotonicity. This is the case when Pi (θ) and θ relationship is determined by the order constraints rule. This assumption means that the probability of individuals to answer the item correctly is measured. It indicates that ability is a monoton non-decreasing function. That is, the individual's measured characteristic the higher the degree of possession the higher the probability of giving the correct answer to the item and / or means to increase. Sijtsma and Molenaar (2002) explained the rule as higher ability individuals are more likely to answer the item correctly. Monotonicity is valid for parametric IRT. The point that distinguishes (non) parametric IRT models from parametric is that ISRF are monotonous but its increase is not based on a certain shape (like logistics etc) (Sijtsma & Molenaar, 2002).

Mokken Scaling, which is based on the NIRT, is one of the commonly used techniques to scale test data. At the same time mokken scale analysis may be use as a dimensionality analysis as well as scale validity (Molenaar & Sijitsma, 2002). It allows examining the dimensionality of the data obtained. In this respect, it allows the examination of the structure of the measured feature and the discovery of its dimensions. Mokken scale analysis is described in two types, namely, explanatory and confirmatory (Mokken, 1971). Confirmatory analysis investigates whether a set of

items are accepted as a scale, or not. In the confirmatory analysis, it is made by defining a certain lower limit for the H coefficient to be estimated for that analysis. “c” indicates the lower bound. In exploratory analysis, items are selected one by one to obtain one or more scales. At each step, the items that will contribute the greatest to the scale are added one by one.

Mokken scaling explains that scale expresses the responses of the respondents as an indicator of a single and latent variable. It presents a nonparametric probabilistic model of the Gutmann scaling approach. It has a probabilistic approach rather than a deterministic approach. In all types and applications of Mokken scaling analysis, H coefficient plays an important role to examine scalability of a scale. H coefficient is associated with Gutmann error (Molenaar, 1997), which gives information when respondent responds correctly to the difficult item but incorrectly answers an easier item. In short, it gives information about unexpected response patterns. When the scale has polytomous items, H coefficients based on Guttman error give the response pattern of respondent who disapproves the more popular step and chooses the less popular one. When the H coefficient is 1, the scale approaches the perfect level according to the Gutmann scaling.

There are three types of H coefficient calculated in Mokken analysis: H coefficient (H_i) for items, item pairs H coefficient (H_{ij}) and H coefficient (H) for scale coefficient is calculated. Theoretically, the H coefficient is expected to take a value between 0 and 1, and all positive H values are acceptable. H = 1 indicates that the Guttman error is approximately equal to 0. “c” value is the lower limit value of H coefficients which can be specified by researcher. It indicates the minimum level of contribution it will make to rank. Therefore, c value indicates information about items to contribute to the scale. It’s suggested that when $c > 0.3$, H coefficients are greater than zero, equal to or greater than c and all the H_{ij} coefficients calculated for item pairs are positive. When the “c” value is determined by the researcher, the scalability coefficient is also controlled. The “c” value can be chosen as 0.3 and above, as well as higher values such as 0.40, 0.50, and stronger scales can be obtained. Benchmark for interpreting H coefficient is suggested by Mokken (1971), $0.3 \leq H < 0.4$ means scale is "weak", $0.4 \leq H < 0.5$ is "medium" and $H \geq 0.5$ is “strong”.

Mokken scale analysis is used to provide evidence for the validity of scores obtained from measurement tools in many areas. It is stated that analysis findings based on non-parametric item response theory are more flexible in assuming the IRT assumptions of the data and providing conditions such as working with a large sample and large item pool. The low number of gifted children and the existence of opinions stating that the concept of giftedness does not have unidimensionality limit the use of parametric IRT models. It is important to benefit from the advantages of IRT, such as not being dependent on items in ability estimation and sampling in parameter estimation, in performing the validity study of the measurement tool for the detection of gifted children, which is based on the examination of ability. It can be said that models based on nonparametric IRT are more suitable for examining the concept of giftedness due to the limitation of sampling, difficulty in assuming normality and unidimensionality.

The aim of this study is to conduct a validity study of the Candidate Notification Scale developed by Bildiren and Bıkmaz Bilgen (2019) to identify gifted children in the Preschool Period with the confirmatory Mokken scaling based on Monotone homogeneity model of non (parametric) item response theory. The study is important in terms of providing more information about the scale with the validation of the framework defined by explanatory and confirmatory factor analysis.

2. Method

In the study, which is descriptive, which is one of the quantitative research methods, it is aimed to collect data by asking questions from the sample in order to define some aspects of the universe (Fraenkel, Wallen & Hyun, 2011). Descriptive research aims to present and interpret the current situation as it is.

2.1 Participants

The study group of the research consists of 253 children (n=122 female, n=131 male) between the ages of 4-6 who attended to 3 preschool education institutions in Aydin. The scale was filled in by the students' own teachers who worked in the pre-school education institution, since it was thought that they knew the students best. The teachers were informed about the identification before filling out the scale and the necessary guidance was provided by the researcher.

2.2 Materials

The "Candidate Notification Scale” developed for gifted children in the Preschool Period", the validity and reliability studies of which were conducted in 2019 by Bildiren and Bıkmaz Bilgen (2019), was used. The scale was filled by

teachers. Each item in the scale consists of thirteen items in the five-point rating scale type scored as 1 (never observed) - 5 (continuously observed). The highest score that can be obtained from the total scale is 65 and the lowest score is 13. In table 1 there is a sample of items from the scale.

Table 1. Sample Items in the Scale

	Never observed (1)	Rarely observed(2)	Sometimes observed(3)	Frequently observed (4)	Always observed (5)
Responds quickly to the questions asked.					
Embarks on original experiments that seem strange to others.					
Works intensely					

The data obtained from the application of the "*Candidate Notification Scale*" were discovered by explanatory factor analysis (Tabachnick & Fidell, 2013), and the data collected from a different sample during the verification of the model were analyzed by confirmatory factor analysis (Jöreskog & Sörbom, 1993). In the reliability study, each sub-dimension of the scale was examined with the Cronbach Alpha reliability coefficient (Crocker & Algina, 1986).

Based on the explanatory factor analysis applied to the data obtained from the application of the scale, the scale consists of 3 dimensions, namely, above average ability, creativity and task commitment, and this three-dimensional structure is verified with $\chi 2$ (62) = 113.46, p = .143, CFI = .99, NNFI = .98, RMSEA = .078, SRMR = .03 values by confirmatory factor analysis. When the reliability is examined on the basis of the dimensions of the scale, *above average ability* has a Cronbach's alpha value of .95, *creativity* has .69 and *task commitment* has .95.

2.3 Data Analysis

In this study, the validity of scale scores have been conducted with Mokken Scaling (monotone homogeneity model) for polytomous items (Molenaar, 1982). Monotone homogeneity model (MHM) was used to realize the Mokken scale analysis, which was developed based on the (non) parametric item response theory, of the scores obtained from the scale. MHM is an (non) parametric IRT model developed by Moleenaar for polytomous items. Parameters estimation with MHM used with the package program: the Mokken Scaling Procedure (Sijtsma, Debets, & Molenaar, 1990). To make estimates with MSP program, all database converted into a format suitable for Mokken scaling analysis

2.3.1 Testing Assumptions

The assumptions of unidimensionality and local independence are related (Lord and Novick, 1968). Two assumptions are assessed together (with automated item selection procedure (AISP)) and at the end of the assessment if unidimensionality is assumed, local independence is assumed too.

In monotonicity testing, researcher examines H coefficients (both coefficient for items (H_i) and for the total scale (H)). If H coefficient is below .40 then the scalability is weak, if the coefficient is between .40-.50 scalability is moderate and if the coefficient is above .50 scalability is strong. Besides we can use "monotonicity assumption testing procedure" which is an procedure in MSP to examine monotonicity. MSP gives "Crit" values which indicates violation of monotonicity when value is above 80 (Molenaar & Sijtsma, 2000). When crit value is under 40, the violation may be because of a sampling error.

On the other hand, Z-values are given for all items of the scale in "short search history" option. For all items MSP check assumption whether they are normally distributed or not. And Z value is estimated for all test items.

2.3.2 Estimation of H Coefficients

The MSP program estimates three different scalability coefficients, H coefficient estimated from all items, H_i coefficient for individual items, H_{ij} coefficient for item-pairs.

1. H coefficient: H coefficient for all items is used to examine model data fit so the scalability of all items can be examined according to the value of this H coefficient. When H=1, it is accepted as a value to express the extent to which a scale items approximates perfect scale. The equation 1 shows the estimation of scalability coefficient H, where Cov indicates the covariance between item pairs and X_i and X_j sembolizes the item scores i and j.

$$[H=\frac{\sum\sum_{i<j} Cov(X_i X_j)}{\sum\sum_{i<j} Cov\ max\ (X_i X_j)} \qquad \textit{Equation 1}$$

Mokken (1971) proposed criteria for the interpretation of the coefficient calculated for the whole scale: If it is between 0.30 and 0.40 it means scale is "weak", if it is between 0.40 and 0.50 it means scale is "medium", if it is over 0.50 it means scale is "strong".

2. H_i coefficient: This coefficient calculated for each item and interpreted as a discrimination index that is accepted as validity criterion (Mokken, 1971). H coefficients calculated with the monotone homogeneity model are called the discrimination coefficient and values close to 1 are interpreted as having a high level of discrimination. Equation 2 shows the estimation formula in which Cov indicates the covariance and X_i and X_j symbolize the item scores i and j.

$$[H_i=\frac{\sum_{i\neq i} Cov(X_iX_j)}{\sum_{i\neq i} Cov \max(X_iX_j)} \qquad \textit{Equation 2}$$

3. H_{ij} coefficient for item-pairs: The H_{ij} calculated for item pairs is a marker of the relationship between items. H coefficient of scalability, H_{ij} can be calculated via Equation 3. It is similar to H and H_i formulas. Cov is covariance, I and j are items:

$$[H_{ij}=\frac{\sum Cov(X_iX_j)}{\sum Cov\ max(X_iX_j)} \qquad \textit{Equation 3}$$

2.3.3 P values

Another value MSP output gives calculated based on MHM is *P(robability)* value. P value is calculated as one less than the number of categories of the item. If item has two categories, only one P value is calculated. When the category of item is five, like in this study, four P values are calculated for each item. The probabilities of item categories are expected to gradually decrease as the category order increases.

2.3.4 Reliability of Scale Scores

Reliability increases as the number of measurement errors decrease. The reliability of scales obtained by MSP is shown by a test-retest procedure similar to Cronbach's alpha, with reliable scales showing test-retest reliability. 0.7 is required for a scale to be considered reliable. Molenaar-Sijstma (MS), Cronbach alpa and lambda reliability coefficient for each scale in determining the reliability of the scale calculated.

3. Results and Discussion

The findings obtained by the Mokken scale analysis of the data obtained from the Candidate Notification Scale developed for Preschool gifted children are given respectively. First automated item selection procedure results are given and then results of testing assumptions and the coefficients of scales estimated are given, respectively.

3.1 Results of Item Selection Procedure

Within the scope of (non)parametric IRT, for unidimensionality in PMTK, another technique used instead of factor analysis applied in Automated Item Selection Procedure (AISP) is the method called. AISP is based on inter-item covariance and uses scalability coefficients (H_i) for items for estimation. MSP has an option to run automated item selection procedure which examines the data in explanatory way. In this procedure the researcher changes the value of lowerbound ("c") to make more homogeneous sub-scales. So, first researcher adds all items to the program (with .30 default value of c) and then changes the "c" value (like .35, .40 etc.) to obtain stronger scales. And researcher can examine detailed "search history" for interpreting results. In detailed history, it is seen that the analysis started with the relationship between item 13 and item12. And in every step an item is added to these items to form a scale. At the end of the steps researcher can decide the scales. In table 2 items giving analysis order is given.

Table 2. Detailed Search History

One Scale			Three Scales	
Item no	Mean	H_i	Item no	Hi
Item 13	3.47	.75	Item 13	.88
Item 12	3.36	.69	Item 12	.85*
Item 11	3.36	.68	Item 11	.87
Item 10	3.15	.73	Item 10	.86
Item 9	3.58	.71	Item 9	.86
Item 2	3.62	.72	Item 2	.86
Item 5	3.34	.73	Item 5	.86
Item 4	3.48	.72	Item 4	.86
Item 3	3.60	.71	Item 3	.86
Item1	3.44	.67	Item1	.84*
Item 8	2.57	.56	Item 8	.53
Item 7	2.54	.56	Item 7	.59
Item 6	3.07	.34*	Item 6	.42*
"c" : .30	r=.95	H=.66	"c" :.30	

When the first AISP analysis is completed, it is seen that all H coefficients (for items) are above zero. The calculated H coefficients vary between .34 and .75 as seen in Table 2. In this structure the fact that the scalability coefficients calculated for all scales are above .30, which is the lower limit for acceptance, indicates that this item can remain in the measurement tool. When the scale is accepted in its current form (unidimensional), it is seen that 12 items are above the value of 0.50. Considering the suggestion of Mokken (1971) in interpreting the coefficient, the scale with a value above this value is considered as a strong scale. The order in which the items are included in the scale is given in Table 2. However, in this construct, it is seen that the 6th item, which is important for the scale, has a low H coefficient (.34) therefore AISP was run one more time by changing the lowerbound value "c" value. And the three scales are described by repeating analysis by investigating H values for items and H value of total scale. It is seen that coefficients is relatively higher in the three dimensional (called three scale) structure. This situation can be expressed in the classical sense that the three-dimensional structure provides better model data fit. For this reason, the results are obtained with the three-dimensional (scale) structure.

3.2 Results of Assumption Testing

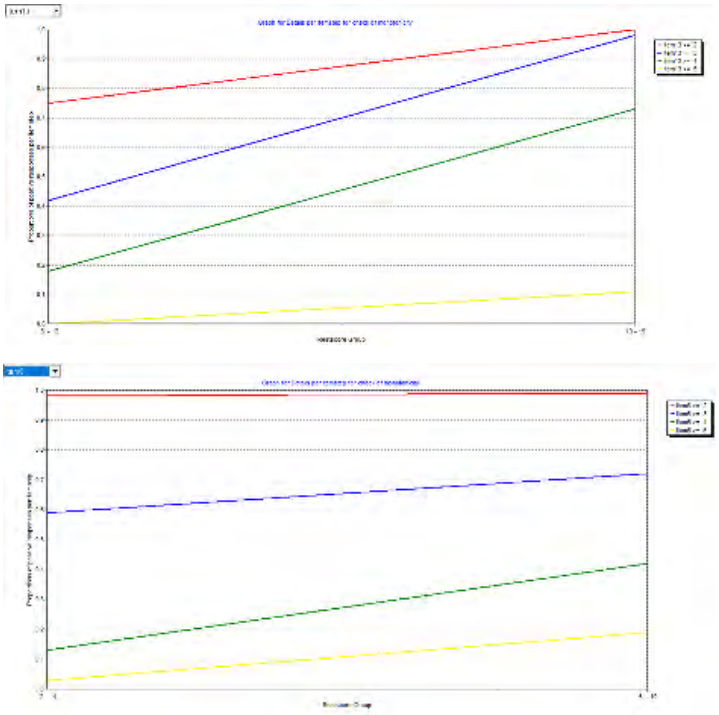
The results are obtained differently for three scales (dimentions). One of the assumptions of IRT models is monotonicity. MSP tests the assumptions by "evaluate model assumptions" procedure. It gives results for monotonicitiy, pmatrix, restcore and restsplit, if these are selected. The values for testing the assumptions are given in Table 3.

Table 3. Examination of Assumptions

Item no	Mean	Monotonicity(Crit)	vi	Zsig	Pmatrx	Restsc.	Z
Item 1	3.42	-	0	-	0	0	24.88
Item 2	3.59	-	0	-	0	0	25.17
Item 3	3.57	-	0	-	0	0	25.41
Item 4	3.45	6*	1	-	0	0	25.84
Item 5	3.59	-	0	-	0	0	24.95
Item 6	3.06	-	0	-	0	0	11.87
Item 7	2.55	22*	1	0	0	0	10.72
Item 8	2.58	12*	2	0	0	0	8.24
Item 9	3.55	-	0	-	0	0	25.43
Item 10	3.13	-	0	-	0	0	25.27
Item 11	3.45	-	0	-	0	0	25.88
Item 12	3.34	-	0	-	0	0	25.49
Item 13	3.44	-	0	-	0	0	26.29

There are 13 items in the scale which have five categories. When the item means estimated by including all the items given in Table 3 are analyzed, it is seen that the item means estimated for the items of five categories are very close to each other and estimated as approximately 3. The low values belong to items 7 and items 8. One of the assumptions of NIRT models is monotonicity. It means that the item step response function is in a structure that does not decrease monotonously, so that individuals with higher ability are more likely to answer the item correctly. (Sijtsma & Molenaar, 2002). The highest violation in the Monotonicity assumption is related to Item 7. But it's not significant based on the criteria of crit value.

In MHM Item step response functions are obtained, and two of them shown in the figure 1. A. ISRF for the item 13 with highest H_i (.88) and B. ISRF item 6 with lowest H_i (.42). Item 13 has moresteeper slopes than Item 6. On the other hand all items assumes the monotonicity with nondecreasing slopes. But it's easily seen that the item step response functions' shapes are different from parametric IRT functions.



A.ISRF (Item 13)

B. ISRF (Item 6)

Figure 1. Graph for Details per Item Step for Check of Monotonicity

*item step shown as red:>=2, blue:>=3, green: :>=4; yellow:>=5

3.3 Results of H Coefficients

The notification scale has three dimentions (Bildiren & Bıkmaz Bilgen, 2019) namely above average ability, creativity and task commitment. In this study confirmation of this model is tested by using mokken scaling analysis. The confirmatory mokken analysis is done by MSP with "test" option. With "test" option, three types of H coefficients are calculated. Two of them (H_i for each item and the scalability coefficient H accepted for the entire scales) are given in Table 4. H_{ij} coefficients for item pairs are shown Table4.

Table 4. H Scalability Coefficients

		Item no	H	(Se)	Z
Scale 1	Above average ability	Item 1	.84*	.03	24.88
		Item 2	.86	.02	25.17
		Item 3	.86	.02	25.41
		Item 4	.86	.02	25.84
		Item 5	.86	.02	24.95
			H=.86	Z_{scale}=39.92	
Scale 2	Creativity	Item 6	.42*	.06*	11.87
		Item 7	.59	.04	10.72
		Item 8	.53	.07	8.24
			H=.51	Z_{scale}=12.54	
Scale 3	Task commitment	Item 9	.86	.02	25.43
		Item 10	.86	.02	25.27
		Item 11	.87	.02	25.88
		Item 12	.85*	.03	25.49
		Item 13	.88	.02	26.29
			H=.86	Z_{scale}=39.68	

Z values were estimated for each scale and each item of the scale. Z values calculated for the whole scales are obtained as 39.92, 12.54, 40.56, respectively. For scale 1 (it corresponds to above average ability dimention) the Z values range from 24.9 to 25.8. The Z values of second scale (dimention of creativity) range from 8.2 to 11.9 and Z values of third scale (dimention of task commitment) range from 25.3 to 26.3 Z values belonging to scale 2 are lower than Z values of the other scales.

While the H scalability coefficient which is estimated by accepting the scale used in the evaluation of model data fit as one-dimensional is .66 (given in Tablo 2), when we consider it in three-dimensional and scale it on the basis of dimensions, the value increases to .86 and .86 for the dimensions of above average ability and task commitment, respectively, however it appears to be lower as .51 for the creativity dimension. The scalability coefficient H estimated for all scales and obtained as .86 for scale1, .51 for scale 2, .86 for scale 3. To accept a scale as the Mokken scale and to say that its assumptions are met, the coefficient H should be above zero. Besides H coefficients above 0.50 indicate strong scales (Mokken, 1971). Findings show that scale 1 and scale 3 are strong scales. On the other hand scale 2 just above the lower limit value for strong scale with a value of 0.51.

When the Hi coefficients (called as discrimination coefficient) for each item are examined for each scale. H value is expected to take a value between 0 and 1 and positive H values are acceptable. If the value of H=1 indicates that Guttman error is approaching 0. Items with a Hi value in the 0.0-0.3 range are positive but low discrimination. It is considered unusable for testing as its contribution to rankings is low. Items with an H value between 0.40 and 0.50 have medium discrimination. H_i values above 0.50 are considered substances with strong discrimination (Mokken, 1971; Sijtsma and Molenear, 2002). When the findings in table 4 are examined, the H coefficients of items belong to scale 1 and scale 3 are over .80, it means all items have strong discrimination. On the other hand the H coefficient varies between .42 and .59 for scale 2. This means the items belong to this scale are medium discrimination. Based on the findings, it is seen that item 6 (with a H value of .42) has a low level of discrimination, while the other 12 items have a high level of discrimination with a value of .50 and above. So it is concluded that all items have enough discrimination power. In Mokken scaling there is one more H coefficient calculated for item pairs, H_{ij}. The Hij coefficients are shown in Table 5.

In table 5, H coefficients for item pairs are given separately for each scale. When MHM is used, there is a requirement that covariances regarding item pairs are not negative (Van der Ark, 2007). This situation is accepted as an indication that model assumptions are met (Sijtsma & Molenaar, 2002). For first scale 10 coefficients estimated, ranged between .78 and .91. For the second scale, 3 coefficients ranging from 0.36 to 0.72 and 10 for the third scale between 0.80 and 0.94 were estimated. When the 23 H_{ij} values estimated for the scales are examined, it is seen that the lowest value belong to item 7 and 8. Based on the finding, the highest H value for item pairs is between item 12 and item 13 with .94. This finding is important in mokken scaling. The AISP process given in Table 2 started the analysis with the highest value item pairs (Item 13 and 12).

Table 5. Matrix of H Values per Item Pair

	Scale 1 Above average ability					Scale 2 Creativity			Scale 3 Task commitment				
Item no	1	2	3	4	5	6	7	8	9	10	11	12	13
1						6			9				
2	.87					7	.72		10	.88			
3	.90	.81				8	.48	.36	11	.85	.90		
4	.82	.87	.86						12	.83	.80	.86	
5	.78	.87	.87	.91					13	.87	.87	.85	.94
	Number of H_{ij} =10 Hij < 0 = 0					Number of H_{ij} =3 Hij < 0 = 0			Number of H_{ij} =10 Hij < 0 = 0				

3.4 Results of Reliability Analysis and P(robability) Values

Within the scope of mokken scale analysis, the accuracy of the item difficulty ranking is also is examined. In this study monotone homogeneity model is used and when P(robability) values are examine, it's seen that the categories the assumption is met. P(robability) values that are given in Table 6. In Table 6, P(robability) values of all items are given according to the individuals' preference for the category. It expresses the probability of individuals to give correct answers to the items. Since each item has five options, a p-value was obtained which was one less than the number of options. For the difficulty (P) values in nonparametric IRT, it is desired that the first value is higher than the second and the second value is higher than the third, respectively.

Table 6. Item P(robabality) Values

Item no	P1	P2	P3	P4	r
Item 1	.98	.83	.47	.13	
Item 2	.98	.88	.60	.12	MS=.96
Item 3	.97	.84	.60	.13	alfa= .95
Item 4	.98	.84	.51	.12	lambda=.95
Item 5	.96	.84	.43	.10	
Item 6	.98	.67	.29	.12	MS=.73
Item 7	.85	.55	.14	.02	alfa= .71
Item 8	.92	.50	.14	.01	lambda=.72
Item 9	.95	.82	.61	.16	
Item 10	.88	.72	.42	.12	MS=.96
Item 11	.96	.81	.53	.14	alfa= .95
Item 12	.94	.78	.48	.13	lambda=.95
Item 13	.93	.81	.58	.12	

Findings examined in the Table 6, all items p values category order are accepted. P_1 values are higher than P_2, P_2 values are higher than P_3 and P_4 values are higher than P_4.

The reliability of the scale was estimated with the MS (Molenaar-Sijtsma) reliability coefficient, alfa and lambda reliability coefficient (Sijtsma & Molenaar, 2002). As this value approaches 1, the reliability of the scores obtained from measurement tool increases. Reliability in the related study was estimated as .96 for scale 1 and scale. On the other hand reliability coefficient for scale to is .73. So it can be say that all scales are high reliability on basis of benchmark for reliability analysis.

4. Conclusion, Discussion and Recommendations

In addition to theoretical studies on the examination of nonparametric IRT models, there is a need for application-based studies. In order to ensure model data fit in cases where it is inevitable to work with a small sample group and an item pool that is not large stated in the literature, it is important in estimations to examine the advantages of the model to be considered related to being more flexible than parametric IRT in meeting assumptions such as unidimensionality, local independence and normality, based on real data (Molenaar, 2001; Sijtsma & Molenaar, 2002; Junker & Sijtsma, 2001).

In the literature, it is stated that if the assumptions of unidimensionality, local independence and normality are met in models based on parametric IRT, the results are more qualified, in other words less erroneous. When the specified assumptions are not met, that is, when the measured quality is not one-dimensional or the sample is small, applied studies are needed to examine the results obtained.

Based on the analysis made with explanatory factor analysis (Tabachnick & Fidell, 2013) and confirmatory factor analysis (Jöreskog & Sörbom, 1993) for the development and confirmation of the scale, a three-dimensional scale was obtained as *above average ability*, *creativity* and *task commitment* (Bildiren & Bıkmaz Bilgen, 2019). In the values estimated based on the nonparametric item response theory, two different findings were obtained regarding the one-factor and three-factor structure of the scale. When model data fit is examined, it is seen that although the scale is considered to be one-dimensional, when the automated item selection procedure scale is handled in three dimensions, the model data fit improves, that is, the scalability coefficient H increases relatively. Reliability increases as the number of measurement errors decrease. When the scale was accepted as one-dimensional, the reliability coefficient was .95, while the values were estimated as .96 for two sub-dimensions and .73 for the other sub-dimension when the scale was considered as three-dimensional.

In cases where the scale is three-dimensional (three scales), it was seen that H coefficient for scale total above the acceptance value for each item and H discrimination coefficient for each item were obtained. Based on this finding, a decision can be made in the dimension structure by examining the theoretical field. The three-dimensional structure obtained displays a structure in harmony with Renzulli's (2000) giftedness, general and special ability level, dedication to work and creativity theory. For model data fit, when the discrimination coefficinets and reliability values of the items are examined, it can be said that the one-dimensional structure of the scale is at an acceptable level, and relatively higher values are obtained for each criterion in the three-dimensional structure compared to the one-dimensional structure. As a result, it can be interpreted that the structure of the scale developed on the basis of confirmatory factor analysis gives a parallel result in the validity study based on nonparametric item response theory.

References

Bildiren, A. (2018). Developmental Characteristics of Gifted Children aged 0-6 years: Parental Observations. *Early Child Development and Care*, *188*(8), 997-1011. https://doi.org/10.1080/03004430.2017.1389919

Bildiren, A., & Bıkmaz Bilgen, Ö. (2019). Okul Öncesi Dönem Üstün Yetenekli Çocuklar için Aday Bildirim Ölçeği: Geçerlik ve Güvenirlik Çalışması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi*, *20*(2), 269-285. https://doi.org/10.21565/ozelegitimdergisi.475278

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Harcourt Brace Jovanovich College Publishers.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. NY: Guilford Press.

Debets, P., & Brouwer, E. (1986). *MSP: a program for Mokken Scale Analysis for Polychotomous Items*. Groningen: iec ProGAMMA.

DeMars, C. (2010). *Item Response Theory*. New York: Oxford University Press. Peer Reviewed Publications.

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

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to Design and Evaluate Research in Education* (8th ed.). New York, NY: The McGraw-Hill Companies

Jöreskog, K., & Sörbom, D. (1993). *LISREL8: Structural Equation Modeling with the SIMPLIS Command language*. Hillsdale, NJ: Erlbaum

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258-272. https://doi.org/10.1177/01466210122032064

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Pub.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement*, *12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage Publications.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis*. Mouton, The Hague.

Mokken, R. J., & Lewis, C. (1982). A Nonparametric Approach to the Analysis of Dichotomous Item Responses. *Applied Psychological Measuremen*t, *6*(4), 417-430. https://doi.org/10.1177/014662168200600404

Molenaar, I. W. (2001). Thirty Years of Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 295-299. https://doi.org/10.1177/01466210122032091

Molenaar, I. W. (1997). Nonparametric Models for Polytomous Responses. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*, 369-380. New York: Springer.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitah'eue Methoden, 3*, 145- 164.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows* [software manual]. Groningen, The Netherlands: iec ProGAMM.

Pfeiffer, S. I. (2015). *Essentials of gifted assessment*. Hoboken, NJ: John Wiley.

Renzulli, J. S. (2000). The Identification and Development of Giftedness as a Paradigm for School Reform. *Journal of Science Education and Technology*, *9*(2), 95-114. https://doi.org/10.1023/a:1009429218821

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken Scaling Analysis for Polychotomous Items: Theory. A Computer Programme and an Empirical Application. *Quality and Quantity, 24*(2), 171-188. https://doi.org/10.1007/bf00209550

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory*. Sage Publications, Thousand Oaks.

Sijtsma, K., & Junker, B. W. (2006). Item Response Theory: Past Performance, Present Developments, and Future Expectations. *Behaviormetrika, 33*(1), 75-102. https://doi.org/10.2333/bhmk.33.75

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson