Ankara University Journal of Faculty of Educational Sciences
Year: 2025, Volume: 58, Issue: 3, 1197-1228
DOI: 10.30964/auebfd.1675814, E-ISSN: 2458-8342, P-ISSN: 1301-3718

Comparison Generalized Item Location Indices for Polytomous Items: A Monte Carlo Simulation Study

ARTICLE TYPE	Received Date	Accepted Date	Published Date
Research Article	04.14.2025	10.17.2025	12.15.2025

Emrah Gül [1]
Hakkari University

Abstract

Generalized Item Location Indices (GILI) is a method that deals with the process of developing or selecting polytomous items based on a single value. This method converts multiple location indices obtained for polytomous items into a single index. The purpose of this research is to examine the performance of GILI under different simulation conditions. For the study, how three different GILI (LI_{mean} -LI_{median}-LI_{IRF}) change according to the number of categories (3, 5 and 7), location parameter (-2, -1, 0, 1 and 2) and sample size (200, 500, and 1000) were compared with a Monte Carlo simulation. According to the results, the LI_{median} was estimated with the highest error in all conditions. On the other hand, LI_{mean} and LI_{IRF} produce similar amounts of error for all conditions. Although LI_{mean} and LI_{IRF} produce similar results at -1, 0, and +1 location levels, LI_{mean} makes more accurate predictions at -2 and +2 location levels. It was concluded that as the number of categories increases, the amount of error calculated in small samples increases. LI_{mean} - LI_{median} - LI_{IRF} values, which are matched with the individual's ability in tests developed for different purposes and CAT applications, can be a good parameter for determining which item to choose next duringtest administration. As a result, the proposed method is being easier and faster will facilitate for practitioners in the item selection process.

Keywords: generalized item location indices, simulation, selecting items, polytomous item

Citation: Gül, E. (2025). Comparison Generalized Item Location Indices for Polytomous Items: A Monte Carlo Simulation Study. *Ankara University Journal of Faculty of Educational Sciences, 58*(3), 1197-1228. https://doi.org/10.30964/auebfd.1675814

[1]*Corresponding Author*: Assist. Prof., Hakkari University, Faculty of Education, Departmant of Educational Sciences, Hakkari, Türkiye, E-mail: emrahgul@hakkari.edu.tr, https://orcid.org/0000-0001-8799-3356, https://ror.org/00nddb461
All ethical declarations related to this article are provided on the final page of the manuscript. (Page: 1228).

The validity and reliability of measurement tools used in education and psychology are directly proportional to the psychometric properties of the items. Well-structured and well-selected items from the item pool reveal the quality of the measurement tool. Psychological constructs that cannot be measured directly, such as achievement, attitude, interest or belief, are measured indirectly through an observed variable. In achievement tests, binary (1-0) categories are often used because they are more informative and reliable (Ali, et al., 2015). It is observed that there is a move away from multiple choice items in the field of measurement and evaluation, including achievement tests (Dodd, De Ayala & Koch, 1995). On the other hand, it is known that multi-category scoring is used to measure latent constructs such as attitude, interest or belief. These are models known as ordered response categories (strongly disagree - strongly agree) that show large differences in measurements between the number of these categories. Within the framework of item response theory, the graded-response model (GRM; Samejima, 1969), the partial credit model (PCM; Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the nominal response model (Bock, 1972) are used to create these measurement tools and to determine item and individual parameters. Polytomous IRT models are less researched than two-category models. This is because the models are more complex and difficult to interpret (Hambleton et al., 2011). The location points obtained from polytomous IRT models are the most important and critical values (Ostini and Nering, 2006). The lack of a specific index representing the position of multiple scored items leads to inadequate selection of such items (Kang, Arbet, Betts & Muntean, 2024). Here, two or more location values are calculated depending on the number of categories. While a single difficulty parameter is calculated for dichotomously scored instruments, more than one difficulty (location) parameter value is calculated for multiple scored instruments. In particular, the development of multiple-scored achievement tests or the use of a single index in CAT applications facilitates item selection. This makes it difficult to evaluate the item. In their report, Ali et al. (2015) made suggestions on how to develop or select multicategorical items based on a single value. For a four-category item, assume that each response is 0, 1, 2 and 3. In this case, the probabilities (θ) of an individual responding to item i will be (θ); $P_{i0}(\theta)$, $P_{i1}(\theta)$, $P_{i2}(\theta)$ and $P_{i3}(\theta)$. In this case, at least four parameter values will be calculated and it will be very difficult to determine a single difficulty or location parameter value for the item due to multiple possibilities. On the other hand, while it is easy to interpret binary scored items because a single item characteristic curve is formed, it becomes very difficult to interpret multiple scored items when separate curves are drawn for each response. One of the item response models that can be used for multiple scored items is the partial credit model (PCM).

Partial Credit Model

The discrimination parameter was not evaluated in the study. Therefore, the Partial Credit Model (PCM) is preferred only when comparing the location parameter value. Another reason for using PCM is its frequent application among ordered models. The Partial Credit Model is a modified form of the Rasch Model (Masters,

1982). Equation 1 below presents the Partial Credit Model. Here, it is assumed that there is a measurement instrument with *k* items. The number of response categories for each item *j* (*j* = 1, 2...,*k*) is m_j. The PCM defines the probability of observing the *l*th (*l* = 0, 1,...,m_j) response category to the *j*th item as a function of the latent trait. δ_{jl} is the location, of each response category. The location value δ_{jl} will be equal to *l*-1 in this case. For example, for an item with 4 categories, a location value of 4-1 = 3 will be calculated.

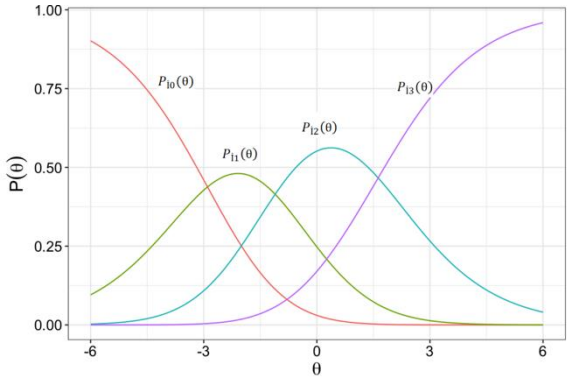
Equation 1

$$P\left(X_{ij}=l|\theta,\boldsymbol{\delta}_j\right)=\frac{\exp\left(l\theta-\sum_{a=1}^{l}\delta_{ja}\right)}{\sum_{b=0}^{m_j}\exp\left(b\theta-\sum_{a=1}^{b}\delta_{ja}\right)} \qquad \theta\sim\mathcal{N}(\mu,\sigma^2)$$

As can be seen in Figure 1, a four-category response model for an item answered with four categories is presented. With the four categories response model, 3 δ_{jl} location values and one discrimination value will be calculated as explained in Equation 1. Since the partial credit model is an extension of the Rasch model, discrimination values are not taken into account. For item i, a_i, b_{i1}, b_{i2} and b_{i3} will be calculated. As mentioned above, since the possible responses are 0, 1, 2 and 3, there are four response categories ($P_{i0}(\theta)$,$P_{i1}(\theta)$, $P_{i2}(\theta)$ *and* $P_{i3}(\theta)$) as shown in Figure 1.

Figure 1

Partial Credit Model



Ali et al. (2015) developed a generalized item location indices model that combines related response categories based on a mathematical formula. Generalized

item location indices calculation methods are based on Item Category Response Function and Item Response Function.

Generalized Item Location Indices Based on the Item Category Response Function

LI_{mean}

The first proposed location index (LI) of item location is the average item category location using all item category response functions (ICRF) and is given in equation 2.

Equation 2

$$LI_{mean} = \frac{1}{m}\sum_{c=1}^{m} b_{ic}$$

LI_{median}

The second index of location is the median of all ICRFs. LI_{median} is given in equation 3.

Equation 3

$$LI_{median} = Median(b_{ix}s) = \begin{cases} b_{ix}^{(k)} \\ 0.5(b_{ix}^{(k)} + b_{ix*}^{(k+1)}) \end{cases}$$

Generalized Item Location Indices Based on the Item Response Function

LI_{IRF}

This form of LI is derived from the polytomous IRF. As is well known, in binary IRF models, such as logistic models with one, two or three parameters, the conditional mean of the item score (expected value) is equal to the probability of answering the item correctly. A location value of 0.5 is used to locate the item with the highest score (for 1-0) of 1. This corresponds to an expected score of 0.5 xm (m being the highest possible category response) in a polytomous IRF (Ali et al., 2015). LI_{IRF} is formulated in equation 4.

Equation 4

$$LI_{IRF} = \theta : E\,[X_i] = \frac{m}{2}$$

When developing measurement instruments, the selection of items to be included in the test is important for the validity and reliability of the test. Test developers use polytomous item types especially when measuring psychological constructs such as attitudes, interests or beliefs. They use polytomous item types not only for the constructs mentioned above but also for mixed achievement tests. Ali

(2011) has been used in many studies that the generally generalized substance position indexes developed (Christensen, et al., 2021; Santoso, Setawati, Ismail, et al., 2023; Talkkari & Rosenström, 2024). In this case, which item to include in the test emerges as an important problem. Not only in the test development process but also in CAT processes, there is a need to develop methods based on a single value for selecting items from the item pool. In this case, using the generalized item indices mentioned above is a good way. Standard indices provide separate calculations for each item category. The indices proposed here, on the other hand, provide a single assessment for each item through separate calculations. Therefore, they facilitate a more efficient item development and selection process.

In this study, the performances of the different indices mentioned above were evaluated under various conditions through simulation. For this purpose, the following questions were addressed.

1. According to the number of different categories in the data structure (3, 5, and 7), location parameter (-2, -1, 0, 1, and 2) and sample size (200, 500 and 1000);
    a. How does LI_{mean} - LI_{median} - LI_{IRF} change?
    b. What is the standard error rates obtained from LI_{mean}-LI_{median}-LI_{IRF} estimates?

Method

In this study, Monte Carlo simulations were conducted to compare generalized item location indices developed for polytomous items (Harwell et al., 1996). In this method, the computer generates data according to probabilistic distributions and allows comparison of the outputs obtained from the model (Sigal & Chalmers, 2016).

Simulation Conditions

Three different generalized item location indices (LI_{mean} - LI_{median} - LI_{IRF}) were compared according to the number of categories (3, 5, and 7), location parameter (-2, -1, 0, 1, and 2) and sample size (200, 500 and 1000). Thus, 3x5x3 = 45 different conditions were obtained and 25 replications were performed for each condition and 1125 data sets were obtained. The simulation conditions are summarized in the table.

Table 1
The Simulation Conditions

Number of Categories	Sample Size	Location Parameter (δ_j)
3	200	-1
		-2
		0
		+1
		+2
5	500	-1
		-2
		0
		+1
		+2
7	1000	-1
		-2
		0
		+1
		+2
Total		Simulation Conditions = 45

While determining the number of categories, Jacoby and Matell (1971) stated that the number of categories does not affect the meaning of the research result and that 3 categories can be used. On the other hand, in a study examining the validity and reliability values of scales consisting of 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11 categories, Preston and Colman (2000) stated that the highest validity level is good for the number of 7-10 categories, moderate for 5-6 categories and low for 2, 3 and 4 (Preston & Colman, 2000). Therefore, in this study, 3, 5, and 7 categories with low, medium and high values were determined as research conditions. When determining the item location level, it is stated that the items to be included in a test have a location level between -2 and +2 (Hambleton et al., 1991). Therefore, to exemplify each location level, the location levels used were -2, -1, 0, 1, and 2 in the study. While determining the sample size, Şahin and Anıl (2017) reported in their study that a sample size of at least 150 would be sufficient for tests of 10, 20, and 30 items in order to calculate the b parameter value correctly. On the other hand, many studies have reported that the

minimum sample size should be 200 (Demars, 2010; Wright and Stone, 1979). Dubravka and Shenghai (2024) revealed that the RMSE values they calculated in their research within the scope of polytomous IRT typically increase at sample sizes of 100 and 250. Therefore, for the purposes of this study, sample sizes of 200, 500 and 1000 were used.

Generation of Data

WinGen software was used to generate the data. The WinGen computer program was developed to generate binary and multi-group item response data for various IRT models and many situations that arise in practice (Han, 2006). A fixed 10-item measurement tool condition was created for all conditions. In data generation using WinGen software, the ability (θ) distribution was first determined. The ability distribution is a normal distribution with mean 0 and standard deviation 1. This approach is identified as the most appropriate method for the real data set in many studies. Therefore, the ability parameters were produced according to the standard normal distribution.

After the ability distribution was determined, the discrimination and location values were determined according to the Partial Credit Model. The discrimination value is determined as a constant (a = 1) according to the model used. The location parameter was kept constant within each simulation condition but was set to different values for other conditions (-2, -1, 0, 1, and 2). We compared how the number of categories (3, 5, and 7) varied according to the location parameter (-2, -1, 0, 1, and 2) and sample size (200, 500 and 1000). Thus, 3x5x3 = 45 different conditions were obtained and 25 repetitions were performed for each condition. According to Harwell, Stone, Hsu, and Kirisci (1996), when the number of repetitions is 25, the errors produced decrease and the effect level approaches 1. Therefore, 25 repetitions were used in this study.

Ethical Committee Approval

This study does not require ethical approval. No human or animal participants were involved, and only simulation data were used.

Data Analysis

R software (R Core Team, 2021) was used for data analysis. R software is an open source program that can analyze data with the help of its built-in packages. The "mirt" package was used in R software. "mirt" provides data analysis using unidimensional and multidimensional item analysis models under Item Response Theory (Chalmers, 2012). The Root Mean Square Error (RMSE) statistic was used to evaluate the errors contained in the generalized item location estimated under conditions where the number of categories (3, 5, and 7), location parameter mean (-2, -1, 0, 1, and 2) and sample sizes (200, 500, and 1000) varied. RMSE, or root mean square error, is the standardized version of the difference between the estimated parameters and the true parameters. RMSE is given in Equation 5.

Equation 5

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\bar{\tau}_{nj} - \tau_{nj})^2}{N}}$$

Results

The averages of Generalized Item Location Indices and PCM parameter values are given in Table 2. For better interpretation of the findings, sample size and location level were taken as reference. On the other hand, the mean values estimated for each category without changing the item order are given; For example, for a category number of 3, the first three items are given; for a category number of 5, the items 4-7 are given; and for a category number of 7, the last 3 items are given.

When Table 1 is examined, as the number of categories increased, small samples (SS = 200) caused larger deviations in the estimation. The data shows that the most accurate predictions are at the '0' location level. On the other hand, the most accurate predictions for location level '0' were calculated when the number of categories was '5'. When the generalized item location indices (LI_{mean} - LI_{median} -LI_{IRF}) were examined, it was calculated that the LI_{median} value caused the highest deviation as the number of categories increased, although it produced similar results in small category numbers. As the sample size increases, it can be inferred that predictions become more accurate in lower categories.

As in Table 2, δ_{jl} = -1 in Table 3 and δ_{jl} = +1 in Table 4 are set for better interpretation of the parameters obtained. When Tables 3 and 4 are examined that the deviations decrease as the sample size increases.

Table 2

Item Parameters (PCM) and the Generalized Item Location Indices (δ_{jl} =-2, 0 and +2)

		PCM parameters						Generalized İtem Location İndices			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-2,970	-1,515	-	-	-	-	-2,243	-2,242	-2,242	**SS = 200** δ_{jl} = -2
2		-2,887	-1,467	-	-	-	-	-2,177	-2,177	-2,176	
3		-2,724	-1,386	-	-	-	-	-2,055	-2,055	-2,055	
4	5	-3.358	-2.629	-2.032	-1.182			-2.300	-2.331	-2.310	
5		-3.608	-2.839	-2.170	-1.271			-2.471	-2.504	-2.483	
6		-3.644	-2.924	-2.289	-1.303			-2.540	-2.606	-2.561	

(continuing)

Table 2 (continued)

7		-3.546	-2.864	-2.181	-1.305			-2.473	-2.522	-2.488	
8	7	-3.273	-2.852	-2.574	-2.248	-1.867	-1.073	-2.314	-2.331	-2.321	
9		-3.389	-2.963	-2.615	-2.256	-1.770	-0.997	-2.411	-2.435	-2.396	
10		-3.410	-2.934	-2.577	-2.216	-1.793	-1.000	-2.348	-2.370	-2.352	
		PCM parameters						Generalized Item Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-0.473	0.570					0.048	0.049	0.048	**SS = 500** $\delta_{jl} = 0$
2		-0.475	0.603					0.063	0.064	0.064	
3		-0.485	0.597					0.055	0.055	0.055	
4	5	-0.797	-0.241	0.184	0.737			-0.029	-0.028	-0.029	
5		-0.802	-0.245	0.204	0.763			-0.019	-0.020	-0.019	
6		-0.807	-0.253	0.191	0.763			-0.026	-0.031	-0.027	
7		-0.800	-0.254	0.180	0.729			-0.036	-0.036	-0.036	
8	7	-0.864	-0.453	-0.170	0.079	0.364	0.787	-0.042	-0.045	-0.043	
9		-0.877	-0.465	-0.171	0.067	0.345	0.766	-0.055	-0.051	-0.055	
10		-0.848	-0.432	-0.158	0.081	0.365	0.792	-0.033	-0.038	-0.034	
		PCM parameters						Generalized Item Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	1.386	2.696					2.041	2.041	2.041	**SS = 1000** $\delta_{jl} = +2$
2		1.404	2.725					2.064	2.064	2.064	
3		1.348	2.638					1.993	1.993	1.993	
4	5	1.163	2.161	2.833	3.510			2.416	2.497	2.441	
5		1.164	2.121	2.769	3.465			2.379	2.444	2.399	
6		1.141	2.095	2.762	3.509			2.376	2.428	2.393	
7		1.132	2.102	2.756	3.411			2.350	2.429	2.374	
8	7	1.047	1.850	2.342	2.724	3.077	3.648	2.447	2.532	2.477	
9		1.013	1.827	2.307	2.678	3.068	3.513	2.401	2.492	2.438	
10		1.008	1.841	2.335	2.726	3.103	3.585	2.433	2.530	2.471	

Table 3

Item Parameters (PCM) and the Generalized Item Location Indices (δ_{jl} =-1)

		PCM parameters						Generalized İtem Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-1.539	-0.432					-0.985	-0.985	-0.985	**SS = 200** δ_{jl} = -1
2		-1.630	-0.416					-1.022	-1.022	-1.022	
3		-1.612	-0.443					-1.027	-1.027	-1.027	
4	5	-1.789	-1.279	-0.815	-0.184			-1.016	-1.046	-1.024	
5		-1.774	-1.234	-0.753	-0.122			-0.971	-0.993	-0.977	
6		-1.754	-1.234	-0.786	-0.157			-0.982	-1.009	-0.990	
7		-1.824	-1.296	-0.810	-0.151			-1.020	-1.053	-1.028	
8	7	-2.210	-1.740	-1.396	-1.125	-0.735	-0.153	-1.226	-1.260	-1.238	
9		-2.195	-1.729	-1.426	-1.119	-0.758	-0.169	-1.232	-1.272	-1.245	
10		-2.195	-1.713	-1.413	-1.107	-0.764	-0.174	-1.227	-1.259	-1.239	

		PCM parameters						Generalized Item Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-1.567	-0.418					-0.992	-0.992	-0.992	**SS = 500** δ_{jl} = -1
2		-1.575	-0.405					-0.990	-0.990	-0.990	
3		-1.602	-0.419					-1.010	-1.010	-1.010	
4	5	-1.846	-1.296	-0.819	-0.152			-1.028	-1.057	-1.036	
5		-1.827	-1.282	-0.810	-0.167			-1.021	-1.046	-1.027	
6		-1.883	-1.315	-0.822	-0.176			-1.049	-1.068	-1.054	
7		-1.853	-1.310	-0.847	-0.185			-1.048	-1.078	-1.056	
8	7	-1.813	-1.425	-1.142	-0.870	-0.551	-0.058	-0.976	-1.005	-0.995	
9		-1.819	-1.425	-1.142	-0.874	-0.560	-0.060	-0.979	-1.007	-1.004	
10		-1.838	-1.426	-1.142	-0.881	-0.565	-0.057	-0.984	-1.011	-0.978	

		PCM parameters						Generalized Item Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-1.535	-0.401					-0.968	-0.968	-0.968	**SS = 1000** δ_{jl} = -1
2		-1.496	-0.391					-0.943	-0.943	-0.943	
3		-1.509	-0.394					-0.951	-0.951	-0.951	
4	5	-1.829	-1.269	-0.799	-0.149			-1.011	-1.033	-1.017	
5		-1.797	-1.249	-0.773	-0.137			-0.989	-1.011	-0.994	
6		-1.799	-1.264	-0.796	-0.150			-1.002	-1.029	-1.009	
7		-1.808	-1.265	-0.783	-0.132			-0.997	-1.024	-1.004	
8	7	-2.018	-1.577	-1.285	-0.999	-0.649	-0.090	-1.103	-1.142	-1.115	
9		-2.054	-1.592	-1.283	-0.999	-0.631	-0.090	-1.108	-1.141	-1.118	
10		-2.035	-1.577	-1.270	-0.981	-0.637	-0.095	-1.099	-1.125	-1.108	

Table 4

Item Parameters (PCM) and the Generalized Item Location Indices ($\delta_{jl} = +1$)

		PCM parameters						Generalized Item Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	0.435	1.550					0.992	0.992	0.992	**SS = 200** $\delta_{jl} = +1$
2		0.504	1.634					1.069	1.069	1.069	
3		0.502	1.652					1.077	1.077	1.077	
4	5	0.211	0.861	1.338	1.921			1.082	1.099	1.087	
5		0.172	0.832	1.364	1.900			1.067	1.098	1.075	
6		0.183	0.865	1.366	1.977			1.097	1.115	1.102	
7		0.199	0.854	1.411	1.918			1.095	1.132	1.104	
8	7	-0.020	0.524	0.905	1.183	1.502	1.980	1.012	1.044	1.021	
9		0.034	0.574	0.928	1.235	1.547	2.047	1.060	1.081	1.066	
10		-0.005	0.550	0.934	1.238	1.538	1.998	1.042	1.085	1.054	
		PCM parameters						Generalized Item Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	0.528	1.694					1.111	1.111	1.111	**SS = 500** $\delta_{jl} = +1$
2		0.512	1.647					1.079	1.079	1.079	
3		0.492	1.618					1.055	1.055	1.055	
4	5	0.149	0.829	1.292	1.850			1.030	1.060	1.038	
5		0.214	0.861	1.364	1.918			1.088	1.112	1.094	
6		0.168	0.850	1.363	1.924			1.076	1.106	1.084	
7		0.191	0.868	1.355	1.915			1.082	1.111	1.089	
8	7	0.047	0.552	0.883	1.138	1.405	1.819	0.974	1.010	0.985	
9		0.051	0.560	0.858	1.135	1.416	1.828	0.974	0.996	0.983	
10		0.055	0.543	0.855	1.127	1.406	1.805	0.964	0.990	0.973	
		PCM parameters						Generalized Item Location			
Item *i*	*NC*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	0.503	1.609					1.056	1.056	1.056	**SS = 1000** $\delta_{jl} = +1$
2		0.503	1.643					1.072	1.072	1.072	
3		0.507	1.646					1.076	1.076	1.076	
4	5	0.160	0.816	1.282	1.820			1.019	1.048	1.027	
5		0.192	0.842	1.300	1.843			1.044	1.070	1.050	
6		0.184	0.815	1.294	1.810			1.025	1.054	1.033	

(continuing)

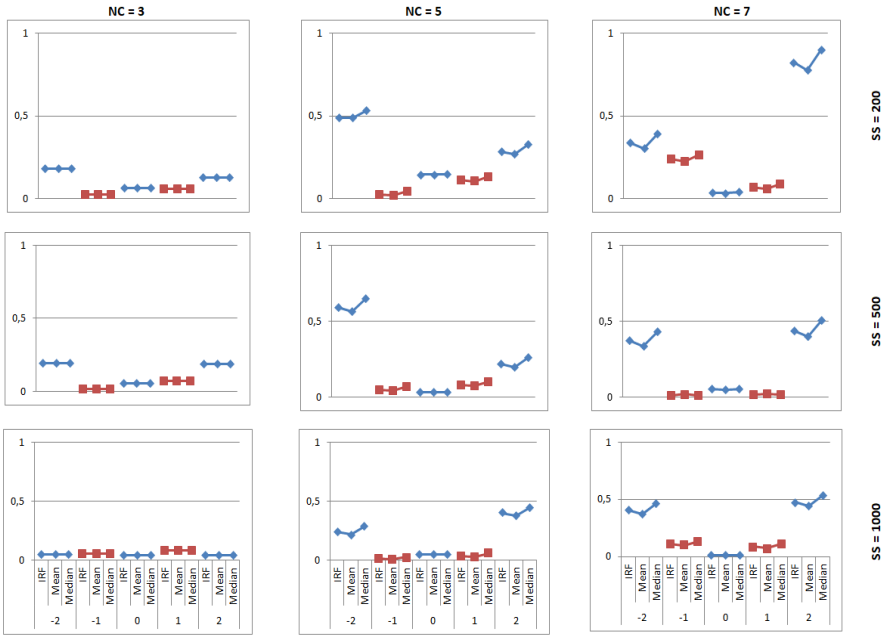
Table 4 (continued)

7		0.165	0.811	1.287	1.824			1.021	1.048	1.029	
8	7	0.073	0.622	0.979	1.252	1.557	1.985	1.077	1.115	1.090	
9		0.054	0.615	0.967	1.246	1.547	1.978	1.067	1.106	1.081	
10		0.075	0.621	0.980	1.264	1.559	1.990	1.081	1.122	1.094	

The highest deviations were found at the δjl = -1 level and at a sample size of 200. In small sample sizes (N = 200), as the number of categories increases, deviations in all location indices are high. In larger sample sizes, the change in the number of categories did not result in significant deviations in the calculated generalized item location indices.

For all conditions, the mean error (RMSE) values for each generalized item location indices (LImean - LImedian - LIIRF) estimation are given in Figure 2. When the Figure is analyzed, it is seen that the error values decrease as the sample size increases in the number of similar categories. On the other hand, it was found that the lowest error rates for all conditions were calculated at -1, 0, and 1 location levels. When the number of categories (NC) is '3', the mean error (RMSE) values for the estimation of generalized location indices (LImean -LImedian-LIIRF) are very close to each other. As the number of categories increases, the generalized item location values also differ. For example, when the number of categories is '5' or '7', the highest amount of error is observed in the 'Median' value. On the other hand, when different location conditions (-2, -1, 0, 1, and 2) were considered, it was observed that the highest amount of error was calculated at the values of -2 and 2. The lowest errors were calculated when the location levels of the items in the test were -1, 0 and +1. Especially in small samples, better predictions were made when the number of categories was low, while the error rate increased as the number of categories increased. Even if the sample size is large, it is seen that the error increases at -2 and +2 location levels

Figure 2
RMSE Values for Generalized Item Location Indices



Discussion, Conclusion and Suggestions

In this study, Monte Carlo simulations were conducted to compare generalized item location indices developed for polytomous items. The performances of the generalized item location indices, estimated under conditions where the number of categories, location parameter, and sample sizes were varied were compared. Based on the results of this research, it can be concluded that including items from the -1, 0, and +1 location ranges in the test, regardless of sample size during item selection for the developed tests, results in less measurement error. Considering the performance of generalized item location indices, it was concluded that the highest errors were found in LImedian calculations. At -1, 0, and +1 location levels, LImean and LIIRF produce similar results. At -2 and +2 location levels, LImean makes more accurate predictions. Kılıç et al., (2023) stated in their literature review that the majority of the studies examined did not include item agreement statistics. This study shows that half of the studies examined in this research include polytomous items and only one study includes item-model selection based on item-model fit, which leads to many errors in item selection. For example, Saepuzaman, Istiyono and Haryanto (2022) used the average difficulty level as a reference when evaluating the scale items in their study. In this study, a total sample size of 264 (small) was used for the four-category scale.

According to the results of this study, it is necessary to use fewer categories for small sample sizes. The item selection process is particularly problematic in the development of polytomous tests. Generally, both GRM and PCM calculate more than one difficulty (location) value in the item selection process. Since this multiple location value does not have a single interpretable value, studies have focused on this point because the level of information provided by the item is used for item selection (Dodd, et al., 1995). In these cases, the generalized item location indices mentioned above can be used. The most important parameters used when creating a test are difficulty (location) and discrimination. The test developer can use these two parameters to develop the target test. When developing tests, the ability level is not limited to level '0' alone. Placement tests and selection tests should have different psychometric properties. In these cases, the question of which method should be used when selecting items for the test emerges. The item selection process is critical not only in test development but also in adaptive testing. LI_{mean} - LI_{median} - LI_{IRF} matched with the individual's ability can be a good parameter for which item will be the next item during the administration of the test. The use of LI_{mean} and LI_{IRF} is particularly more effective in item selection and low error estimations according to this research. Haag, et al. (2020) used the LI_{IRF} value instead of location values in their research. Two main reasons were put forward for using the LI_{IRF} index. The first of these is that it is based on a single index instead of four location values, and the other is that the resulting location index focuses on the latent variable itself rather than the responses given to the latent variable. In his study, Ali (2011) found that the use of generalized item location indices for item selection resulted in more effective use of the item pool compared to other methods. Another result of this study is that the proposed method is easier and faster than other methods. As in any simulation study, the conditions of this study may be limited. Different conditions can be considered when developing tests in education and psychology. In this study, ideal conditions were considered. Therefore, studies can be conducted using different locations for discriminations. On the other hand, for this research, a normal distribution with a mean of 0 and a standard deviation of 1 was used. This is because the data obtained are compatible with the real data structure. Therefore, it can be explored how these indexes perform across different talent distributions. Studies addressing this topic are quite limited in the existing literature. Although some research has theoretically discussed the potential advantages of the indexes used, empirical studies that demonstrate the underlying reasons in practice remain scarce (Christensen, Comins, Krogsgaard et al., 2021; Kang, Arbet, Betts & Muntean, 2024; Santoso, Setawati, Ismail et al., 2023; Talkkari & Rosenström, 2024). Therefore, it is recommended that the subject be investigated under different conditions. For example, the effect on results in different ability distributions is an important issue. On the other hand, studies to be conducted with other poly-IRT models will contribute to the field.

Ankara University Journal of Faculty of Educational Sciences
Year: 2025, Volume: 58, Issue: 3, 1197-1228
DOI: 10.30964/auebfd.1677195, E-ISSN: 2458-8342, P-ISSN: 1301-3718

Çok Kategorili Maddeler İçin Genelleştirilmiş Madde Konum İndekslerinin Karşılaştırılması: Monte Carlo Simülasyon Çalışması

MAKALE TÜRÜ	Başvuru Tarihi	Kabul Tarihi	Yayın Tarihi
Araştırma Makalesi	14.04.2025	17.10.2025	15.12.2025

Emrah Gül[1]
Hakkari Üniversitesi

Öz

Genelleştirilmiş Madde Konum İndeksleri (GILI), çok kategorili maddelerin geliştirilmesi veya seçilmesi sürecine tek bir değer üzerinden yaklaşan bir yöntemdir. Bu yöntem, çok kategorili maddeler için elde edilen çeşitli konum indekslerini tek bir indekse dönüştürmektedir. Bu çalışmanın amacı, GILI yönteminin farklı simülasyon koşulları altındaki performansını incelemektir. Araştırmada, üç farklı GILI türünün (LI_{mean} - LI_{median} - LI_{IRF}), kategori sayısı (3, 5 ve 7), konum parametresi (-2, -1, 0, 1 ve 2) ve örneklem büyüklüğü (200, 500 ve 1000) değişkenlerine göre nasıl değiştiği Monte Carlo simülasyonları aracılığıyla karşılaştırmalı olarak analiz edilmiştir. Elde edilen bulgulara göre, tüm koşullarda en yüksek tahmin hatası LI_{median} için gözlemlenmiştir. Buna karşın, LI_{mean} ve LI_{IRF}, tüm koşullarda benzer hata düzeyleri üretmiştir. Özellikle -1, 0 ve +1 konum düzeylerinde benzer sonuçlar üretmelerine rağmen, LI_{mean}'in -2 ve +2 konum düzeylerinde daha isabetli tahminlerde bulunduğu tespit edilmiştir. Ayrıca, kategori sayısının artması durumunda, küçük örneklemlerde hesaplanan hata miktarının da arttığı sonucuna ulaşılmıştır. Farklı amaçlarla geliştirilen testler ve Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) uygulamalarında, bireylerin yetenek düzeyleriyle eşleştirilen LI_{mean} , LI_{median} ve LI_{IRF} değerleri, test sürecinde bir sonraki seçilecek maddenin belirlenmesinde etkili bir parametre olarak kullanılabilir. Sonuç olarak, önerilen yöntemin daha kolay ve hızlı uygulanabilir olması, madde seçimi sürecinde uygulayıcılara önemli ölçüde kolaylık sağlayacaktır.

Keywords: genelleştirilmiş madde konum indeksleri, simülasyon, madde seçme, çok kategorili maddeler.

[1]*Sorumlu Yazar:* Dr. Öğr. Üye., Hakkari Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Hakkari, Türkiye, E-mail: emrahgul@hakkari.edu.tr, https://orcid.org/0000-0001-8799-3356, https://ror.org/00nddb461
Makaleye ilişkin tüm etik beyanlar, çalışmanın son sayfasında yer almaktadır. (Sayfa:1228)

Eğitim ve psikolojide kullanılan ölçme araçlarının geçerliği ve güvenirliği, doğrudan maddeye ilişkin psikometrik özelliklerle ilişkilidir. Madde havuzundan iyi yapılandırılmış ve uygun şekilde seçilmiş maddeler, ölçme aracının kalitesini ortaya koyar. Başarı, tutum, ilgi veya inanç gibi doğrudan ölçülemeyen psikolojik yapılar, dolaylı olarak gözlenen değişkenler aracılığıyla ölçülür. Başarı testlerinde, bilgi verici ve güvenilir olmaları nedeniyle genellikle ikili (1-0) kategoriler kullanılmaktadır (Ali vd., 2015). Ölçme ve değerlendirme alanında, başarı testleri de dahil olmak üzere, çoktan seçmeli maddelerden uzaklaşıldığı gözlemlenmektedir (Dodd vd., 1995). Öte yandan, tutum, ilgi veya inanç gibi örtük (latent) yapıları ölçmek amacıyla çok kategorili puanlamaların kullanıldığı bilinmektedir. Bu puanlamalar, sıralı tepki kategorileri (tamamen katılmıyorum - tamamen katılıyorum) olarak bilinir ve kategori sayısına bağlı olarak ölçüm sonuçlarında büyük farklılıklar göstermektedir. Madde tepki kuramı (Item Response Theory) çerçevesinde; derecelendirilmiş tepki modeli (Graded Response Model, GRM; Samejima, 1969), kısmi kredi modeli (Partial Credit Model, PCM; Masters, 1982), genelleştirilmiş kısmi kredi modeli (Generalized Partial Credit Model, GPCM; Muraki, 1992) ve nominal tepki modeli (Nominal Response Model; Bock, 1972) gibi modeller, bu tür ölçme araçlarının geliştirilmesi ve madde ile birey parametrelerinin belirlenmesinde kullanılmaktadır. Çok kategorili (polytomous) IRT modelleri, ikili (dichotomous) modellere kıyasla daha az araştırılmıştır; çünkü bu modeller daha karmaşık olup yorumlanmaları da zordur (Hambleton ve diğ., 2011). Çok kategorili IRT modellerinden elde edilen konum noktaları, maddenin en önemli ve kritik değerleri arasında yer almaktadır (Ostini ve Nering, 2006). Çok kategorili puanlanan maddelerin konumunu temsil eden özel bir indeksin eksikliği, bu tür maddelerin yetersiz seçilmesine neden olmaktadır (Kang vd., 2024). Bu tür maddelerde, kategori sayısına bağlı olarak iki veya daha fazla konum değeri hesaplanmaktadır. İkili puanlanan maddeler için tek bir güçlük parametresi hesaplanırken, çok kategorili maddeler için birden fazla güçlük (veya konum) parametresi değeri hesaplanmaktadır. Özellikle çok kategorili başarı testlerinin geliştirilmesinde veya Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) uygulamalarında tek bir indeksin kullanılması madde seçimini kolaylaştırmaktadır. Ancak bu durum maddenin değerlendirilmesini zorlaştırmaktadır. Ali vd. (2015), raporlarında çok kategorili maddelerin tek bir değer üzerinden nasıl geliştirilebileceği veya seçilebileceği konusunda önerilerde bulunmuşlardır. Dört kategorili bir madde düşünüldüğünde, her bir yanıtın 0, 1, 2 ve 3 olarak kodlandığı varsayılmaktadır. Bu durumda, bir bireyin madde i'ye θ yetenek düzeyinde yanıt verme olasılıkları; $P_{i0}(\theta)$, $P_{i1}(\theta)$, $P_{i2}(\theta)$ ve $P_{i3}(\theta)$ şeklinde ifade edilir. Bu durumda, en az dört parametre değeri hesaplanacak ve olasılıkların çokluğu nedeniyle maddeye ilişkin tek bir güçlük veya konum parametresinin belirlenmesi oldukça güç hale gelecektir. Öte yandan, ikili puanlanan maddeler için tek bir madde karakteristik eğrisi çizildiğinden yorumlama kolayken, çok kategorili maddelerde her bir yanıt için ayrı eğriler çizilmesi yorumlamayı zorlaştırmaktadır. Çok kategorili maddeler için kullanılabilecek madde tepki modellerinden biri, kısmi kredi modelidir (PCM).

Kısmi Kredi Modeli

Bu çalışmada ayırt edicilik (discrimination) parametresi değerlendirilmemiştir. Bu nedenle, Kısmi Kredi Modeli'nin (Partial Credit Model - PCM) tercih edilmesinin temel gerekçesi yalnızca konum (location) parametresi değerlerinin karşılaştırılmasıdır. PCM'nin kullanılma nedenlerinden bir diğeri ise, sıralı modeller arasında en sık başvurulan modellerden biri olmasıdır. Kısmi Kredi Modeli, Rasch Modeli'nin değiştirilmiş bir formudur (Masters, 1982). Aşağıda Eşitlik 1 de Kısmi Puan Modeli sunulmaktadır. Burada *k* madde sayısı kadar bir ölçme aracının olduğu varsayılmaktadır. Her bir madde *j* (*j* = 1, 2…,*k*)'ye verilen yanıt kategori sayısı *mj* olmaktadır. PCM, gizil özelliğin bir fonksiyonu olarak *j*'inci maddeye verilen (*l* = 0, 1,...,*mj*) yanıt kategorisinin gözlemlenme olasılığını tanımlamaktadır. δ*jl* ise her bir yanıt kategorisine ait güçlük yani yer parametre değeridir. δ*jl* yer değeri bu durumda *l*-1 kadar olacaktır. Örneğin 4 kategorili bir madde için 4-1 = 3 yer parametre değeri hesaplanacaktır.

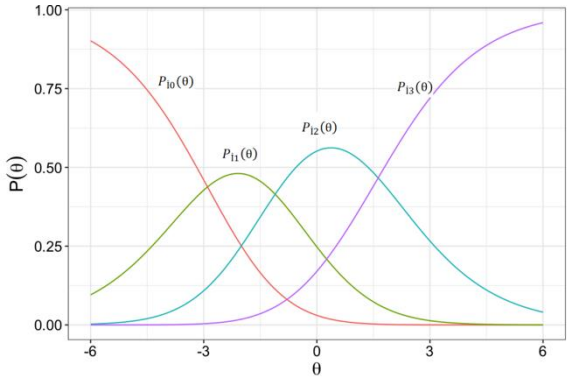
Eşitlik 1

$$P\left(X_{ij}=l|\theta,\boldsymbol{\delta}_j\right)=\frac{\exp\left(l\theta-\sum_{a=1}^{l}\delta_{ja}\right)}{\sum_{b=0}^{m_j}\exp\left(b\theta-\sum_{a=1}^{b}\delta_{ja}\right)}\qquad \theta\sim\mathcal{N}(\mu,\sigma^2)$$

Şekil 1'de görüldüğü üzere, dört kategoride yanıtlanan bir maddeye ilişkin dört kategorili yanıt modeli sunulmuştur. Dört kategorili yanıt modeliyle, Denklem 1'de açıklandığı şekilde üç konum değeri ve bir ayırt edicilik değeri hesaplanmaktadır. Ancak, kısmi puanlama modeli Rasch modelinin bir uzantısı olduğundan, ayırt edicilik değerleri dikkate alınmamaktadır. *i* maddesi için a_i, b_{i1}, b_{i2} and b_{i3} değerleri hesaplanacaktır. Yukarıda belirtildiği üzere, olası yanıtlar 0, 1, 2 ve 3 olduğundan, Şekil 1'de gösterildiği gibi dört yanıt kategorisi bulunmaktadır ($P_{i0}(\theta)$, $P_{i1}(\theta), P_{i2}(\theta)\ and\ P_{i3}(\theta)$).

Şekil 1

Kısmi Kredi Modeli



Ali ve ark. (2015), ilişkili yanıt kategorilerini matematiksel bir formüle dayalı olarak birleştiren genelleştirilmiş bir madde konum indeksleri modeli geliştirmiştir. Genelleştirilmiş madde konum indekslerinin hesaplama yöntemleri, Madde Kategori Yanıt Fonksiyonu ve Madde Yanıt Fonksiyonu'na dayanmaktadır.

Madde Kategori Yanıt Fonksiyonuna Dayalı Genelleştirilmiş Madde Konum İndeksleri

LI_{mean}

Madde konumuna ilişkin önerilen ilk konum indeksi (LI), tüm madde kategori yanıt fonksiyonları (ICRF) kullanılarak hesaplanan ortalama madde kategori konumudur ve Eşitlik 2'de verilmiştir.

Eşitlik 2

$$LI_{mean} = \frac{1}{m}\sum_{c=1}^{m} b_{ic}$$

LI_{median}

İkinci konum indeksi ise tüm madde kategori yanıt fonksiyonlarının (ICRF) medyanıdır. LI_{median}, eşitlik 3'te verilmiştir.

Eşitlik 3

$$LI_{median} = Median(b_{ix}s) = \begin{cases} b_{ix}^{(k)} \\ 0.5(b_{ix}^{(k)} + b_{ix*}^{(k+1)}) \end{cases}$$

Madde Yanıt Fonksiyonuna Dayalı Genelleştirilmiş Madde Konum İndeksleri

LI_{IRF}

Bu LI biçimi, çok ketegorili madde yanıt fonksiyonu'ndan (IRF) türetilmiştir. Bilindiği üzere, ikili IRF modellerinde örneğin bir, iki veya üç parametreli lojistik modellerde madde puanının koşullu ortalaması (beklenen değer), maddeye doğru yanıt verme olasılığına eşittir. En yüksek puanın 1 olduğu (1-0) durumlarda maddeyi konumlandırmak için 0.5 konum parametre değeri kullanılır. Bu, çok ketegorili bir IRF'te beklenen puanın 0.5×m'ye (m en yüksek olası kategori yanıtı olmak üzere) karşılık gelmektedir (Ali ve ark., 2015). LI_{IRF} eşitlik 4'te formüle edilmiştir.

Eşitlik 4

$$LI_{IRF} = \theta : E\,[X_i] = \frac{m}{2}$$

Ölçme araçları geliştirilirken teste dâhil edilecek maddelerin seçimi testin geçerlik ve güvenirliği açısından önemlidir. Test geliştiriciler özellikle tutumlar, ilgiler veya inançlar gibi psikolojik yapıları ölçerken çok kategorili madde tiplerini kullanırlar. Çok kategorili madde tiplerini yalnızca yukarıda belirtilen yapılar için değil, aynı zamanda karma başarı testleri için de kullanırlar. Ali (2011) rapor ettiği üzere birçok çalışmada genelleştirilmiş madde konumu indekslerini kullanmıştır (Talkkari ve Rosenström, 2024; Christensen, Comins, Krogsgaard vd. 2021; Santoso, Setawati, Ismail vd. 2023). Bu durumda teste hangi maddenin dâhil edileceği önemli bir sorun olarak ortaya çıkar. Yalnızca test geliştirme sürecinde değil, Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) süreçlerinde de madde havuzundan madde seçimi için tek bir değere dayalı yöntemler geliştirmeye ihtiyaç vardır. Bu durumda yukarıda belirtilen genelleştirilmiş madde indekslerini kullanmak iyi bir yoldur. Standart indeksler her bir madde kategorisi için ayrı hesaplamalar sağlar. Burada önerilen endeksler ise, her bir madde için ayrı hesaplamalar yoluyla tek bir değerlendirme sunmakta ve bu sayede daha verimli bir madde geliştirme ve seçme sürecini kolaylaştırmaktadır.

Bu çalışmada, yukarıda belirtilen farklı endekslerin performansları simülasyon yoluyla farklı koşullar altında değerlendirilmiştir. Bu amaçla aşağıdaki sorulara yanıt aranmıştır.

Veri yapısındaki farklı kategori sayısına (3, 5 ve 7), konum parametresine (-2, -1, 0, 1 ve 2) ve örneklem büyüklüğüne (200, 500 ve 1000) göre;

a. LI_{mean}-LI_{median}-LI_{IRF} değerleri nasıl değişmektedir?

b. LImean-LImedian-LIIRF kestirimlerinden elde edilen standart hatalar nasıldır?

Yöntem

Bu çalışmada, çok kategorili maddeler için geliştirilen Genelleştirilmiş Madde Konum İndeksleri karşılaştırmak amacıyla Monte Carlo simülasyonları

gerçekleştirilmiştir (Harwell vd., 1996). Bu yöntemde, bilgisayar olasılıksal dağılımlara göre veri üretmekte ve modelden elde edilen çıktıların karşılaştırılmasına olanak sağlamaktadır (Sigal ve Chalmers, 2016).

Simülasyon Koşulları

Üç farklı genelleştirilmiş madde konum indeksi (LI_{mean} - LI_{median} - LI_{IRF}), kategori sayısı (3, 5 ve 7), konum parametresi (-2, -1, 0, 1 ve 2) ve örneklem büyüklüğü (200, 500 ve 1000) değişkenlerine göre karşılaştırılmıştır. Böylece, 3x5x3 = 45 farklı koşul elde edilmiş ve her koşul için 25 yineleme gerçekleştirilerek toplamda 1125 veri seti oluşturulmuştur. Simülasyon koşulları tablo 1'de özetlenmiştir

Tablo 1.
Simülasyon Koşulları

Kategori Sayısı	Örneklem Büyüklüğü	Yer Parametresi (δ_j)
3	200	-1
		-2
		0
		+1
		+2
5	500	-1
		-2
		0
		+1
		+2
7	1000	-1
		-2
		0
		+1
		+2
Toplam		Simülasyon Koşulları= 45

Kategori sayısının belirlenmesi aşamasında, Jacoby ve Matell (1971), kategori sayısının araştırma sonucunun anlamını etkilemediğini ve 3 kategorinin kullanılabileceğini belirtmiştir. Öte yandan, 2, 3, 4, 5, 6, 7, 8, 9, 10 ve 11 kategoriden oluşan ölçeklerin geçerlik ve güvenirlik değerlerini inceleyen bir çalışmada ise Preston ve Colman (2000), en yüksek geçerlik düzeyinin 7-10 kategori için iyi, 5-6 kategori için orta ve 2, 3 ve 4 kategoriler için düşük olduğunu ifade etmiştir (Preston ve Colman, 2000). Bu nedenle, bu çalışmada düşük, orta ve yüksek değerleri temsil eden 3, 5 ve 7 kategori araştırma koşulları olarak belirlenmiştir. Madde konum düzeyi belirlenirken, bir teste dahil edilecek maddelerin konum düzeyinin -2 ile +2 arasında

olduğu belirtilmiştir (Hambleton ve ark., 1991). Bu nedenle, çalışmada her konum düzeyini örneklendirmek amacıyla -2, -1, 0, 1 ve 2 konum düzeyleri kullanılmıştır.

Örneklem büyüklüğünün belirlenmesinde, Şahin ve Anıl (2017), b parametre değerinin doğru hesaplanabilmesi için 10, 20 ve 30 maddelik testler için en az 150 kişilik bir örneklem büyüklüğünün yeterli olduğunu raporlamıştır. Öte yandan, birçok çalışma minimum örneklem büyüklüğünün 200 olması gerektiğini bildirmiştir (Demars, 2010; Wright ve Stone, 1979). Dubravka ve Shenghai (2024) ise polytomous IRT kapsamında yaptıkları araştırmada hesapladıkları RMSE değerlerinin genellikle 100 ve 250 örneklem büyüklüğünün altında arttığını ortaya koymuştur. Bu nedenle, bu çalışmanın amaçları doğrultusunda 200, 500 ve 1000 örneklem büyüklükleri kullanılmıştır.

Verilerin Üretilmesi

Veri oluşturmak için WinGen yazılımı kullanılmıştır. WinGen bilgisayar programı, çeşitli IRT modelleri ve uygulamada karşılaşılan birçok durum için ikili ve çok kategorili madde tepki modellerinde veriler üretmek amacıyla geliştirilmiştir (Han, 2006). Tüm koşullar için sabit 10 maddelik bir ölçme aracı durumu oluşturulmuştur. WinGen yazılımında veri üretiminde öncelikle yetenek (θ) dağılımı belirlenmiştir. Yetenek dağılımı, ortalaması 0 ve standart sapması 1 olan normal dağılım olarak tanımlanmıştır. Bu durum, birçok çalışmada gerçek veri seti için en uygun yöntem olarak kabul edilmektedir. Bu nedenle, yetenek parametreleri standart normal dağılıma göre üretilmiştir. Yetenek dağılımı belirlendikten sonra, Kısmi Kredi Modeline (Partial Credit Model) göre ayrıştırma ve konum değerleri belirlenmiştir. Ayırt edicilik değeri, kullanılan modele göre sabit (a = 1) olarak belirlenmiştir. Konum parametresi her simülasyon koşulu için sabit tutulmuş, diğer koşullar için ise (-2, -1, 0, 1 ve 2) olarak değiştirilmiştir. Kategori sayısının (3, 5 ve 7) konum parametresine (-2, -1, 0, 1 ve 2) ve örneklem büyüklüğüne (200, 500 ve 1000) göre nasıl değiştiği karşılaştırılmıştır. Böylece, 3x5x3 = 45 farklı koşul elde edilmiş ve her koşul için 25 tekrar yapılmıştır. Harwell, Stone, Hsu ve Kirisci (1996) çalışmalarında, tekrar sayısının 25 olduğunda ortaya çıkan hata oranlarının azaldığını ve etki düzeyinin 1'e yaklaştığını tespit etmişlerdir. Bu nedenle, bu çalışmada 25 tekrar kullanılmıştır.

Etik Kurul Kararı

Bu çalışma, etik kurul izni gerektirmemektedir. Araştırmada herhangi bir insan veya hayvan katılımcıdan veri toplanmamış, yalnızca simülasyon verisi kullanılmıştır.

Verilerin Analizi

Veri analizinde R yazılımı (R Core Team, 2021) kullanılmıştır. R yazılımı, içerisinde yer alan paketler aracılığıyla veri analizi yapabilen açık kaynaklı bir programdır. Bu çalışmada, R yazılımı içerisinde "mirt" paketi kullanılmıştır. mirt paketi, Madde Tepki Kuramı (MTK) kapsamında tek boyutlu ve çok boyutlu madde analiz modelleriyle veri analizi yapılmasına olanak sağlamaktadır (Chalmers, 2012). Genelleştirilmiş madde konumunun tahmininde, kategori sayısı (3, 5 ve 7), konum

parametresinin ortalaması (-2, -1, 0, 1 ve 2) ve örneklem büyüklüğü (200, 500 ve 1000) gibi koşulların değiştiği durumlarda ortaya çıkan hataların değerlendirilmesinde Kök Ortalama Kare Hata (Root Mean Square Error - RMSE) istatistiği kullanılmıştır. RMSE, tahmin edilen parametreler ile gerçek parametreler arasındaki farkların standartlaştırılmış bir göstergesi ya da kök ortalama kare hatasıdır. RMSE, Denklem 5'te verilmiştir.

Eşitlik 5

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\bar{\tau}_{nj} - \tau_{nj})^2}{N}}$$

Bulgular

Genelleştirilmiş Madde Konum İndeksleri ile Kısmi Kredi Modeli (PCM) parametre değerlerinin ortalamaları Tablo 2'de verilmiştir. Bulguların daha iyi yorumlanabilmesi amacıyla örneklem büyüklüğü ve konum düzeyi referans olarak alınmıştır. Diğer yandan, madde sırası değiştirilmeden her bir kategori sayısı için tahmin edilen ortalama değerler sunulmuştur. Örneğin, kategori sayısı 3 olan durum için ilk üç madde, kategori sayısı 5 olan durum için 4 ila 7. maddeler ve kategori sayısı 7 olan durum için son üç madde dikkate alınmıştır. Tablo 1 incelendiğinde, kategori sayısı arttıkça küçük örneklem (SS = 200) koşullarında tahminlerdeki sapmaların büyüdüğü görülmektedir. En doğru tahminlerin '0' konum düzeyinde elde edildiği anlaşılmaktadır. Diğer yandan, konum düzeyi '0' için en doğru tahminlerin kategori sayısı '5' olduğunda yapıldığı hesaplanmıştır. Genelleştirilmiş madde konum indeksleri (LI_{mean} - LI_{median} - LI_{IRF}) incelendiğinde, küçük kategori sayılarında benzer sonuçlar üretse de, kategori sayısı arttıkça LImedian değerinin en yüksek sapmaya neden olduğu hesaplanmıştır. Örneklem büyüklüğü arttıkça, düşük kategori sayılarında daha doğru tahminler yapıldığı sonucuna ulaşılmıştır. Tablo 2'de olduğu gibi, elde edilen parametrelerin daha iyi yorumlanabilmesi amacıyla Tablo 3'te δjl = -1 ve Tablo 4'te δjl = +1 olarak alınmıştır. Tablo 3 ve 4 incelendiğinde, örneklem büyüklüğü arttıkça sapmaların azaldığı görülmektedir.

Tablo 2

Madde Parametreleri (PCM) ve Genelleştirilmiş Madde Konum İndeksleri (δ_{jl} =-2, 0 and +2)

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-2,970	-1,515	-	-	-	-	-2,243	-2,242	-2,242	**ÖB = 200** δ_{jl} = -2
2		-2,887	-1,467	-	-	-	-	-2,177	-2,177	-2,176	
3		-2,724	-1,386	-	-	-	-	-2,055	-2,055	-2,055	
4	5	-3.358	-2.629	-2.032	-1.182			-2.300	-2.331	-2.310	
5		-3.608	-2.839	-2.170	-1.271			-2.471	-2.504	-2.483	
6		-3.644	-2.924	-2.289	-1.303			-2.540	-2.606	-2.561	
7		-3.546	-2.864	-2.181	-1.305			-2.473	-2.522	-2.488	
8	7	-3.273	-2.852	-2.574	-2.248	-1.867	-1.073	-2.314	-2.331	-2.321	
9		-3.389	-2.963	-2.615	-2.256	-1.770	-0.997	-2.411	-2.435	-2.396	
10		-3.410	-2.934	-2.577	-2.216	-1.793	-1.000	-2.348	-2.370	-2.352	

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-0.473	0.570					0.048	0.049	0.048	**ÖB = 500** δ_{jl} = 0
2		-0.475	0.603					0.063	0.064	0.064	
3		-0.485	0.597					0.055	0.055	0.055	
4	5	-0.797	-0.241	0.184	0.737			-0.029	-0.028	-0.029	
5		-0.802	-0.245	0.204	0.763			-0.019	-0.020	-0.019	
6		-0.807	-0.253	0.191	0.763			-0.026	-0.031	-0.027	
7		-0.800	-0.254	0.180	0.729			-0.036	-0.036	-0.036	
8	7	-0.864	-0.453	-0.170	0.079	0.364	0.787	-0.042	-0.045	-0.043	
9		-0.877	-0.465	-0.171	0.067	0.345	0.766	-0.055	-0.051	-0.055	
10		-0.848	-0.432	-0.158	0.081	0.365	0.792	-0.033	-0.038	-0.034	

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	1.386	2.696					2.041	2.041	2.041	**ÖB = 1000** δ_{jl} = +2
2		1.404	2.725					2.064	2.064	2.064	
3		1.348	2.638					1.993	1.993	1.993	
4	5	1.163	2.161	2.833	3.510			2.416	2.497	2.441	
5		1.164	2.121	2.769	3.465			2.379	2.444	2.399	
6		1.141	2.095	2.762	3.509			2.376	2.428	2.393	
7		1.132	2.102	2.756	3.411			2.350	2.429	2.374	
8	7	1.047	1.850	2.342	2.724	3.077	3.648	2.447	2.532	2.477	
9		1.013	1.827	2.307	2.678	3.068	3.513	2.401	2.492	2.438	
10		1.008	1.841	2.335	2.726	3.103	3.585	2.433	2.530	2.471	

Tablo 3

Madde Parametreleri (PCM) ve Genelleştirilmiş Madde Konum İndeksleri (δ_{jl} =-1)

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-1.539	-0.432					-0.985	-0.985	-0.985	**ÖB = 200** δ_{jl} = *-1*
2		-1.630	-0.416					-1.022	-1.022	-1.022	
3		-1.612	-0.443					-1.027	-1.027	-1.027	
4	5	-1.789	-1.279	-0.815	-0.184			-1.016	-1.046	-1.024	
5		-1.774	-1.234	-0.753	-0.122			-0.971	-0.993	-0.977	
6		-1.754	-1.234	-0.786	-0.157			-0.982	-1.009	-0.990	
7		-1.824	-1.296	-0.810	-0.151			-1.020	-1.053	-1.028	
8	7	-2.210	-1.740	-1.396	-1.125	-0.735	-0.153	-1.226	-1.260	-1.238	
9		-2.195	-1.729	-1.426	-1.119	-0.758	-0.169	-1.232	-1.272	-1.245	
10		-2.195	-1.713	-1.413	-1.107	-0.764	-0.174	-1.227	-1.259	-1.239	

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-1.567	-0.418					-0.992	-0.992	-0.992	**ÖB = 500** δ_{jl} = *-1*
2		-1.575	-0.405					-0.990	-0.990	-0.990	
3		-1.602	-0.419					-1.010	-1.010	-1.010	
4	5	-1.846	-1.296	-0.819	-0.152			-1.028	-1.057	-1.036	
5		-1.827	-1.282	-0.810	-0.167			-1.021	-1.046	-1.027	
6		-1.883	-1.315	-0.822	-0.176			-1.049	-1.068	-1.054	
7		-1.853	-1.310	-0.847	-0.185			-1.048	-1.078	-1.056	
8	7	-1.813	-1.425	-1.142	-0.870	-0.551	-0.058	-0.976	-1.005	-0.995	
9		-1.819	-1.425	-1.142	-0.874	-0.560	-0.060	-0.979	-1.007	-1.004	
10		-1.838	-1.426	-1.142	-0.881	-0.565	-0.057	-0.984	-1.011	-0.978	

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	-1.535	-0.401					-0.968	-0.968	-0.968	**ÖB = 1000** δ_{jl} = *-1*
2		-1.496	-0.391					-0.943	-0.943	-0.943	
3		-1.509	-0.394					-0.951	-0.951	-0.951	
4	5	-1.829	-1.269	-0.799	-0.149			-1.011	-1.033	-1.017	
5		-1.797	-1.249	-0.773	-0.137			-0.989	-1.011	-0.994	
6		-1.799	-1.264	-0.796	-0.150			-1.002	-1.029	-1.009	
7		-1.808	-1.265	-0.783	-0.132			-0.997	-1.024	-1.004	
8	7	-2.018	-1.577	-1.285	-0.999	-0.649	-0.090	-1.103	-1.142	-1.115	
9		-2.054	-1.592	-1.283	-0.999	-0.631	-0.090	-1.108	-1.141	-1.118	
10		-2.035	-1.577	-1.270	-0.981	-0.637	-0.095	-1.099	-1.125	-1.108	

Tablo 4

Madde Parametreleri (PCM) ve Genelleştirilmiş Madde Konum İndeksleri ($\delta_{jl} = +1$)

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	0.435	1.550					0.992	0.992	0.992	**ÖB = 200** $\delta_{jl} = +1$
2		0.504	1.634					1.069	1.069	1.069	
3		0.502	1.652					1.077	1.077	1.077	
4	5	0.211	0.861	1.338	1.921			1.082	1.099	1.087	
5		0.172	0.832	1.364	1.900			1.067	1.098	1.075	
6		0.183	0.865	1.366	1.977			1.097	1.115	1.102	
7		0.199	0.854	1.411	1.918			1.095	1.132	1.104	
8	7	-0.020	0.524	0.905	1.183	1.502	1.980	1.012	1.044	1.021	
9		0.034	0.574	0.928	1.235	1.547	2.047	1.060	1.081	1.066	
10		-0.005	0.550	0.934	1.238	1.538	1.998	1.042	1.085	1.054	

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	0.528	1.694					1.111	1.111	1.111	**ÖB = 500** $\delta_{jl} = +1$
2		0.512	1.647					1.079	1.079	1.079	
3		0.492	1.618					1.055	1.055	1.055	
4	5	0.149	0.829	1.292	1.850			1.030	1.060	1.038	
5		0.214	0.861	1.364	1.918			1.088	1.112	1.094	
6		0.168	0.850	1.363	1.924			1.076	1.106	1.084	
7		0.191	0.868	1.355	1.915			1.082	1.111	1.089	
8	7	0.047	0.552	0.883	1.138	1.405	1.819	0.974	1.010	0.985	
9		0.051	0.560	0.858	1.135	1.416	1.828	0.974	0.996	0.983	
10		0.055	0.543	0.855	1.127	1.406	1.805	0.964	0.990	0.973	

		PCM parametreleri						Genelleştirilmiş Madde Konum İndeksleri			
Madde *i*	*KS*	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	LI_{mean}	LI_{median}	LI_{IRF}	
1	3	0.503	1.609					1.056	1.056	1.056	**ÖB = 1000** $\delta_{jl} = +1$
2		0.503	1.643					1.072	1.072	1.072	
3		0.507	1.646					1.076	1.076	1.076	
4	5	0.160	0.816	1.282	1.820			1.019	1.048	1.027	
5		0.192	0.842	1.300	1.843			1.044	1.070	1.050	
6		0.184	0.815	1.294	1.810			1.025	1.054	1.033	

(devam ediyor)

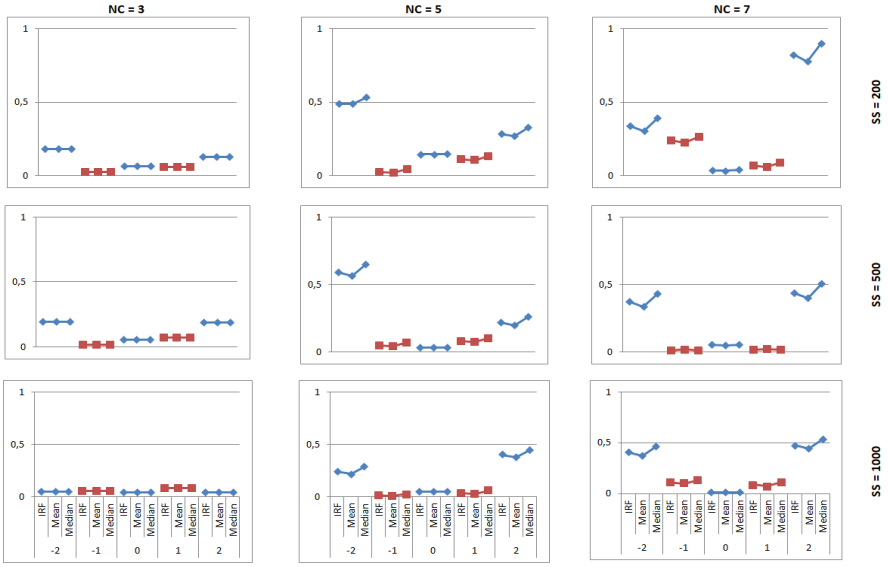
Tablo 4 (devamı)

7		0.165	0.811	1.287	1.824			1.021	1.048	1.029	
8	7	0.073	0.622	0.979	1.252	1.557	1.985	1.077	1.115	1.090	
9		0.054	0.615	0.967	1.246	1.547	1.978	1.067	1.106	1.081	
10		0.075	0.621	0.980	1.264	1.559	1.990	1.081	1.122	1.094	

En yüksek sapmalar, δjl = -1 düzeyinde ve 200 kişilik örneklem büyüklüğünde gözlemlenmiştir. Küçük örneklemlerde (N = 200), kategori sayısı arttıkça tüm konum indekslerinde sapmaların yüksek olduğu görülmektedir. Öte yandan, büyük örneklemlerde kategori sayısındaki değişim, hesaplanan genelleştirilmiş madde konum indeksleri üzerinde büyük sapmalara yol açmamıştır. Tüm koşullar için, her bir genelleştirilmiş madde konum indeksinin (LI_{mean} - LI_{median} - LI_{IRF}) tahminine ilişkin ortalama hata (RMSE) değerleri Şekil 2'de sunulmuştur. Şekil incelendiğinde, benzer kategori sayılarına sahip durumlarda örneklem büyüklüğü arttıkça hata değerlerinin azaldığı görülmektedir. Diğer yandan, tüm koşullar için en düşük hata oranlarının -1, 0 ve 1 konum düzeylerinde hesaplandığı bulunmuştur. Kategori sayısı (KS) 3 olduğunda, genelleştirilmiş konum indekslerinin (LI_{mean} - LI_{median} - LI_{IRF}) tahminine ilişkin ortalama hata (RMSE) değerleri birbirine oldukça yakındır. Ancak kategori sayısı arttıkça, genelleştirilmiş madde konum değerleri arasında farklılık oluşmaktadır. Örneğin, kategori sayısı 5 veya 7 olduğunda, en yüksek hata "Median" değerinde gözlemlenmektedir. Farklı konum koşulları (-2, -1, 0, 1 ve 2) ele alındığında ise en yüksek hata miktarının -2 ve 2 değerlerinde hesaplandığı görülmektedir. En düşük hata değerleri, testteki maddelerin konum düzeylerinin -1, 0 ve +1 olduğu durumlarda hesaplanmıştır. Özellikle küçük örneklemlerde, kategori sayısı düşük olduğunda daha doğru tahminler elde edilirken, kategori sayısı arttıkça hata oranı da artmaktadır. Örneklem büyüklüğü büyük olsa dahi, -2 ve +2 konum düzeylerinde hata oranlarının arttığı gözlemlenmektedir.

Şekil 2

Genelleştirilmiş Madde Konum İndeksleriler ait RMSE Değerleri



Tartışma, Sonuç ve Öneriler

Bu çalışmada, çok kategorili maddeler için geliştirilen genelleştirilmiş madde konum indekslerini karşılaştırmak amacıyla Monte Carlo simülasyonları gerçekleştirilmiştir. Kategori sayısı, konum parametresi ve örneklem büyüklüğünün değiştirildiği koşullar altında tahmin edilen genelleştirilmiş madde konum indekslerinin performansları karşılaştırılmıştır. Elde edilen sonuçlar incelendiğinde, geliştirilecek testler için madde seçimi yapılırken örneklem büyüklüğünden bağımsız olarak -1, 0 ve +1 konum aralıklarında yer alan maddelerin teste dâhil edilmesi durumunda daha düşük hata elde edileceği sonucuna ulaşılmıştır. Genelleştirilmiş madde konum indekslerinin performansı dikkate alındığında, en yüksek hataların LImedian hesaplamalarında ortaya çıktığı görülmüştür. -1, 0 ve +1 konum düzeylerinde LImean ve LIIRF benzer sonuçlar üretirken, -2 ve +2 konum düzeylerinde LImean daha doğru tahminler yapmaktadır. Kılıç vd. (2023) literatür taramasında, incelenen çalışmaların çoğunda madde uyum istatistiklerine yer verilmediğini belirtmiştir. Bu çalışma ise, incelenen çalışmaların yarısında çok kategorili maddelere yer verildiğini ve yalnızca bir çalışmada madde-model uyumuna dayalı madde-model seçimine yer verildiğini göstermektedir. Bu durum, madde seçiminde çok sayıda hataya yol açmaktadır. Örneğin, Saepuzaman, Istiyono ve Haryanto (2022) çalışmalarında ölçek maddelerini değerlendirirken ortalama zorluk düzeyini referans almışlardır. Dört kategorili ölçek için toplam 264 kişilik (küçük) örneklem büyüklüğü kullanılmıştır. Bu çalışmanın sonuçlarına göre, küçük örneklem

büyüklüklerinde daha az kategori sayısına sahip ölçeklerin kullanılması önerilmektedir. Çok kategorili testlerin geliştirilmesinde madde seçimi süreci oldukça sorunludur. Hem GRM (Dereceli Tepki Modeli) hem de PCM (Kısmi Kredi Modeli) madde seçimi sürecinde birden fazla zorluk (konum) değeri hesaplamaktadır. Bu çoklu konum değerlerinin tek bir yorumlanabilir değeri bulunmadığından, çalışmalarda genellikle maddenin sağladığı bilgi düzeyi esas alınarak madde seçimine gidilmektedir (Dodd ve diğ., 1995). Bu tür durumlarda, yukarıda bahsedilen genelleştirilmiş madde konum indeksleri kullanılabilir. Test oluşturulurken dikkate alınması gereken en önemli parametreler zorluk (konum) ve ayırt ediciliktir. Test geliştiriciler bu iki parametreyi kullanarak hedef testlerini oluşturabilirler. Ancak, test geliştirirken yetenek düzeyi yalnızca "0" düzeyinde değildir. Yerleştirme ve seçme testlerinin farklı psikometrik özelliklere sahip olması gerekmektedir. Bu gibi durumlarda, testte yer alacak maddelerin seçilmesinde hangi yöntemin kullanılacağı sorusu ortaya çıkmaktadır. Madde seçimi süreci yalnızca test geliştirme sürecinde değil, aynı zamanda bilgisayar ortamında bireye uyarlanmış testlerde (BOBUT) de oldukça önemlidir. LImean, LImedian ve LIIRF, bireyin yeteneğiyle eşleştirildiğinde test sırasında bir sonraki sorunun seçilmesinde etkili bir parametre olabilir. Özellikle bu çalışmanın sonucuna göre LImean ve LIIRF'in madde seçiminde daha etkili olduğu ve daha düşük hata tahminleri sağladığı görülmektedir. Haag vd. (2020) çalışmalarında konum değerleri yerine LIIRF değerini kullanmışlardır. LIIRF indeksinin kullanımında iki temel gerekçe öne sürülmüştür: İlki, dört ayrı konum değeri yerine tek bir indeksin esas alınması; ikincisi ise, ortaya çıkan konum indeksinin gizil değişkene verilen yanıtlardan ziyade gizil değişkenin kendisine odaklanmasıdır. Ali (2011) çalışmasında, madde seçiminde genelleştirilmiş madde konum indekslerinin kullanılmasının diğer yöntemlere kıyasla madde havuzunun daha etkili kullanımına olanak sağladığını bulmuştur. Bu çalışmanın bir diğer sonucu da önerilen yöntemin daha kolay ve hızlı olmasıdır. Her simülasyon çalışmasında olduğu gibi, bu çalışmada da ele alınan koşullar sınırlı olabilir. Eğitim ve psikoloji alanlarında test geliştirme sürecinde farklı koşullar dikkate alınabilir. Bu çalışmada ideal koşullar varsayılmıştır. Bu nedenle, farklı ayırt edicilik düzeylerine sahip konum değerleriyle çalışılabilecek araştırmalar yürütülebilir. Öte yandan, bu araştırmada ortalamaları 0 ve standart sapmaları 1 olan normal dağılım kullanılmıştır. Bunun nedeni, elde edilen verilerin gerçek veri yapısıyla uyumlu olmasıdır. Bu nedenle, söz konusu indekslerin farklı yetenek dağılımlarında nasıl çalıştığı da incelenebilir. Bu konuyla ilgili literatürdeki çalışmalar oldukça sınırlıdır. Her ne kadar bazı çalışmalar bu indekslerin olası yararlarını teorik olarak ortaya koymuş olsa da, pratikte nedenlerini açıklayan çalışmalar sınırlıdır (Christensen, Comins, Krogsgaard vd., 2021; Kang vd., 2024; Santoso, Setawati, Ismail vd., 2023; Talkkari ve Rosenström, 2024). Bu nedenle, konunun farklı koşullar altında araştırılması önerilmektedir. Örneğin, farklı yetenek dağılımlarında sonuçların nasıl etkileneceği önemli bir araştırma alanıdır. Öte yandan, diğer çok kategorili MTK (polytomous IRT) modelleriyle yapılacak çalışmalar da alana katkı sağlayacaktır.

References

Ali, U. S. (2011). Item selection methods in polytomous computerized adaptive testing (Unpublished doctoral dissertation). University of Illinois, Urbana, IL.

Ali, U. S., Chang, H.-H., & Anderson, C. J. (2015). Location indices for ordinal polytomous items based on itemresponse theory (Research Report No. RR-15-20). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12065

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 27, 29–51. http://hdl.handle.net/10.1007/BF02291411

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Christensen KB, Comins JD, Krogsgaard MR, et al.(2021) Psychometric validation of PROM instruments. Scand J Med Sci Sport.; 31: 1225–1238. https://doi.org/10.1111/sms.13908

DeMars, C. (2010) Item Response Theory: Understanding Statistics Measurement. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. Applied Psychological Measurement, 19(1), 5-22. https://doi.org/10.1177/014662169501900103

Dodd, B., Gıllon, G., Oerlemans, M., Russell, T., Syrmıs, M. & Wılson, H., (1995), Phonological disorder and the acquisition of literacy. In B. Dodd (ed.), Differential Diagnosis and Treatment of Children with Speech Disorder (London: Whurr), pp. 125–146.

Dubravka S. V. & Shenghai D. (2024) Number of Response Categories and Sample Size Requirements in Polytomous IRT Models, The Journal of Experimental Education, 92:1, 154-185, DOI: 10.1080/00220973.2022.2153783

Hambleton RK, Swaminathan H, & Rogers HJ. (1991). Fundamentals of Item Response Theory. Thousand Oaks

Hambleton RK, van der Linden WJ, & Wells CS. (2011). IRT models for the analysis of polytomous scored data: Brief and selected history of model building advances. In: Nering ML, Ostini R, editors. Handbook of Polytomous Item Response Theory Models. p. 21–42

Han, K.T. (2006). WinGen: Windows software that generates IRT parameters and item responses [Computer software]. Amherst, MA: Center for Educational Assessment, University of Massachusetts. http://people.umass.edu/kha/wingen/

Haag DG, Santiago PHR, Macedo DM, Bastos JL, Paradies Y, Jamieson L (2020) Development and initial psychometric assessment of the race-related attitudes and multiculturalism scale in Australia. PLoS ONE 15(4): e0230724. https://doi.org/10.1371/journal.pone.0230724

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. Applied Psychological Measurement, 20(2), 101–125. https://doi.org/10.1177/014662169602000201

Jacoby, J. & Matell, M.S. (1971), Three-point likert scales are good enough, Journal of Marketing Research, 9(4), 444-446. DOI:10.1177/002224377100800414

Kang, H.-A., Arbet, G., Betts, J., & Muntean, W. (2024). Location-Matching Adaptive Testing for Polytomous Technology-Enhanced Items. Applied Psychological Measurement, 48(1-2), 57-76. https://doi.org/10.1177/01466216241227548

Kılıç, A. F., Koyuncu, İ, & Uysal, İ. (2023). Scale development based on item response theory: A systematic review. International Journal of Psychology and Educational Studies, 10(1), 209-223. https://dx.doi.org/10.52380/ijpes.2023.10.1.982

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159–176. https://doi.org/10.1177/014662169201600206

Ostini, R., & Nering, M. L. (2006). Polytomous item response theory models. SAGE.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. Acta Psychologica, 104(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org

Saepuzaman, D., Edi Istiyono, & Haryanto. (2022). Characteristics of fundamental physics higher-order thinking skills test using Item Response Theory analysis. Pegem Journal of Education and Instruction, 12(4), 269–279. https://doi.org/10.47750/pegegog.12.04.28

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. Educational Sciences: Theory & Practice, 17, 321–335. http://dx.doi.org/10.12738/estp.2017.1.0270

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometrika MonographNo. 17). Richmond, VA: Psychometric Society.

Santoso, P.H., Setiawati, F.A., Ismail, R. et al. (2023). Comparing IRT properties among different category numbers: a case from attitudinal measurement on physics education research. Discov Psychol 3, 39. https://doi.org/10.1007/s44202-023-00101-6

Sigal, M. J. & Chalmers, R. P. (2016). Play it again: Teaching statistics with monte carlo simulation. Journal of Statistics Education, 24(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Talkkari, A. & Rosenström, T. H. (2024). Non-Gaussian Liability Distribution for Depression in the General Population. Assessment, 0(0). https://doi.org/10.1177/10731911241275327

Wright, B. D., & Stone, M. H. (1979). Best Test Design: Rasch Measurement. Chicago, IL: Mesa Press.

Ethical and Author Declarations | Etik ve Yazar Beyanları

Authors' Contributions

This study was conducted by a single author. All stages of the research (conceptualization, data collection, analysis, interpretation, and writing) were carried out by the author.

Yazarların Katkı Düzeyleri

Bu çalışma tek yazar tarafından yürütülmüştür. Çalışmanın tüm aşamaları (konu seçimi, veri toplama, analiz, yorumlama ve yazım) yazar tarafından gerçekleştirilmiştir.

Conflict of Interest

The author report there is no competing interests to declare.

Çıkar Çatışması Beyanı

Yazar, beyan edilecek herhangi bir çıkar çatışmasının olmadığını bildirmektedir.

Ethical Approve

This study does not require ethical approval. No human or animal participants were involved, and only simulation data were used.

Etik Onay

Bu çalışma, etik kurul izni gerektirmemektedir. Araştırmada herhangi bir insan veya hayvan katılımcıdan veri toplanmamış, yalnızca simülasyon verisi kullanılmıştır.

Use of Artificial Intelligence

During the preparation of this work the author used CHATGPT 5.0 in order to language improvement. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Yapay Zeka Kullanımı

Yazar, bu çalışmanın hazırlanması sırasında dil geliştirme amacıyla CHATGPT 5.0'ı kullanmıştır. Bu aracı/hizmeti kullandıktan sonra, yazar içeriği gerektiği gibi gözden geçirip düzenlemiş ve yayınlanan makalenin içeriğinin tüm sorumluluğunu üstlenmiştir.

This study has been evaluated under double-blind peer review and verified to be free of plagiarism using iThenticate software.
Bu çalışma, çift taraflı kör hakemlik kapsamında değerlendirilmiş ve iThenticate yazılımı kullanılarak intihal içermediği teyit edilmiştir.

Ethical disclosure | Etik bildirim: ebfd@ankara.edu.tr