International Journal of Assessment Tools in Education
2022, Vol. 9, No. 3, 697–712
https://doi.org/10.21449/ijate.1059628
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A Comparison of type I error and power rates in procedures used determining test dimensionality

Gul Guler[1,*], **Rahime Nukhet Cikrikci**[1]

[1]Istanbul Aydın University, Faculty of Education, Department of Elementary Education, Türkiye

ARTICLE HISTORY
Received: Jan. 18, 2022
Revised: June 30, 2022
Accepted: Aug. 14, 2022

Keywords:
Dimensionality,
Construct validity,
DIMTEST T statistic,
DETECT,
Nonlinear factor analysis.

Abstract: The purpose of this study was to investigate the Type I Error findings and power rates of the methods used to determine dimensionality in unidimensional and bidimensional psychological constructs for various conditions (characteristic of the distribution, sample size, length of the test, and interdimensional correlation) and to examine the joint effect of the conditions (effect of the interaction of conditions) as well as the main effect of each condition. The simulative data were generated for the study using the SAS program. Within the scope of the study, the data were analyzed using the DIMTEST T statistic and the Dimensionality DETECT IDN index, which is one of the non-parametric methods. The Nonlinear Factor Analysis (NOHARM) method was preferred from among parametric methods. As a result of the study, it was noted that the most consistent results in making the unidimensionality decisions belong to the Nonlinear Factor Analysis method showing standard normal distribution according to the shape of the distribution. When the power study results were examined, it was noted that the DIMTEST T statistic gave more accurate results in conditions with large samples, consisting of data with standard normal distribution. On the other hand, while results of the DETECT IDN index and Nonlinear factor analysis were more internally consistent, it was noted that in conditions where the sample size was 1000 and above, the DIMTEST T statistic also made the right decisions in determining dimensionality.

1. INTRODUCTION

In the process of test and scale development in education and psychology, dimensionality is frequently used in validity studies. Dimensionality is the relationship between the items in a test and the implicit feature that the test is thought to measure (Svetina, 2011). Dimensionality is related to the number of skills or psychological constructs that a test or item set measures. The dimensionality determination process is an important issue to consider, regardless of whether the measurement model is unidimensional or multidimensional (Embretson & Reise, 2000). A test has a theoretical structure and is prepared for a specific purpose. The underlying structure of the test should be examined and verified. In this context, construct validity studies are important in terms of the technical features of instruments in education and psychology and

*CONTACT: Gul Guler gulyuce2010@gmail.com Istanbul Aydın University, Faculty of Education, Department of Elementary Education, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

are one of the necessary steps in assessing the dimensionality of tests and scales. A feature to be measured may be associated with more than one implicit feature by nature. When we look at the tests used in education and psychology, it is seen that most of them measure more than one latent feature. For example, while a science test was developed to measure science process skills, it could also measure reading comprehension. For this reason, it is useful to know whether the structure to be measured is one-dimensional or multidimensional. Considering the purpose of creating and applying the test, this situation will affect the validity of decisions made about individuals based on test scores. Determining the dimensionality of the items in a test is extremely important as it will also shape the statistical analysis of the data (Svetina, 2011; Zhang, 2008).

In case a measurement procedure is treated as unidimensional while being in fact multidimensional, the interpretation of test scores, and thus the validity of measurement processes would be misleading (Göçer Şahin, 2016; Touron et al., 2012). Determination of dimensionality, in addition to the determination of the extent to which unidimensionality is neglected and revealing the power of tests with Type Iorsignificant in terms of the validity of decisions made as a result of the tests applied. When a test is unidimensional, that is, when the H***o*** hypothesis is true, accepting the H1 hypothesis with a statistical decision, meaning that the test is multidimensional, causes a Type I Error. Accepting the H***o*** hypothesis while a test is multidimensional, in other words, saying it is unidimensional causes a Type II Error. In addition, deciding that a test is statistically multidimensional while it is actually multidimensional displays the power of the test. Thus, it is considered that testing of unidimensionality is required since the determination of all these situations is directly related to the validity of the decisions.

When studies in the literature are assessed, dimensionality determination methods are generally separated as parametric and non-parametric methods (Abswoude et al., 2004; Mroch & Bolt, 2006; Özbek, 2012; Reinchenberg, 2013; Svetina, 2011; Svetina & Levy, 2014). Conditions such as small samples, low numbers of items, and a high degree of interdimensional correlation revealed the need to study and use non-parametric methods and comparison conditions in addition to parametric methods. The purpose of this study is to investigate the Type I Error and power rates of the methods used to determine dimensionality in unidimensional and two-dimensional psychological constructs depending on sample size, characteristics of the distribution, test length, and interdimensional correlation conditions while comparing the main effect of each condition in addition to joint effects of conditions (effect of the interaction of conditions). In line with this general purpose, answers were sought to the following questions:

1. How do *Type I Error rates* obtained from unidimensional data change where the length of the test, characteristics of distribution, and sample size are manipulated, according to various dimensionality determination methods, in tests scored dichotomously?

2. How do *power rates of the test*, obtained from bidimensional data change where test length, interdimensional correlation degree, distributions and sample size are manipulated according to various dimensionality determination methods, in tests scored dichotomously?

3. What are the Type I Error rates and the power rates of the test using standard, normal and skewed data according to various dimensionality determination methods in tests scored dichotomously?

The most significant reason for choosing the DIMTEST T statistic in this study was the fact that it was a testing method that worked well in large samples and large item pools, and it was effective in displaying even small secondary features (Svetina, 2011). The reason for preferring Nonlinear Factor Analysis was that its results could be interpreted easily, it worked well in small samples, and it was based on factor analytical approaches. In addition, the fact that all methods were accessible for free supported the preference (Svetina & Levy, 2014; Touron et

al., 2012). While factor analysis is generally preferred in unidimensional studies, many studies stated that examining unidimensionality with factor analysis alone is not sufficient and recommended other methods (Finch & Monahan, 2008; Hattie et al., 1996; Ledesma & Valero-Mora, 2007; Özbek, 2012; Reichenberg 2013; Svetina, 2011; Svetina & Levy, 2014; Touron et al., 2012; Yen, 2007). Despite this argument, in many national or international studies factor analysis is used and considered sufficient in the examination of unidimensionality. However, factor analysis requires the assumption of a multivariate normal distribution which might not be achieved in social sciences frequently.

Applying factor analysis to prove unidimensionality – due to the nature of test and scale development – or not using any methods and calculating test scores over the test totals to arrive at decisions about individuals taking achievement tests at national or international test centers are limiting factors in terms of the validity of the decisions.

The fact that achievement tests used by national or international test centers that use factor analysis only or do not use any methods to accept unidimensionality and calculate test scores over the total test to arrive at decisions about individuals – due to the nature of test and scale development – is a limiting factor in terms of the validity of the decisions. If there is a violation of unidimensionality, the multidimensional structure must be determined with correct methods and indices, and it should be investigated for construct validity studies. Another important point in the process of determining unidimensionality is the requirement for test developers to investigate the effect of sample size on determining unidimensionality considering the difficulties experienced in data collection processes in our country.

2. METHOD

2.1. Data Production Study

In this study, simulation data were used to respond to the research questions. Simulation models should be based on realistic parameters (Davey et al., 1997; as cited in Göçer Şahin, 2016). In addition, simulation studies are meaningful when they are similar to real situations. Since it is difficult to meet all the conditions stated in this study in real data at the same time, it was decided to use simulation data. The data of this study were produced using the SAS software. The data were generated in a 2-parameter logistic and compensatory model for power analysis, in accordance with a dichotomous bidimensional structure. For the Type I Error study, unidimensional dichotomous data was generated in the 2-parameter logistic model. Variables, number of conditions, and condition values are presented in Table 1:

Table 1. *Variables and their conditions used in data production.*

Study	Variables	Number of Conditions	Condition Values
Type I Error study	Properties of Distribution	2	Normal, Skewed
	Sample Size	6	200, 300, 500, 1000, 2000, and 3000
	Test Length	3	10, 20, 30
Power Analysis	Properties of Distribution	2	Normal, Skewed
	Sample Size	6	200, 300, 500, 1000, 2000, and 3000
	Test Length	3	10, 20, 30
	Interdimensional correlation	4	0.25, 0.50, 0.75, 0.90
Number of Replications		100	

When Table 1 is examined, considering the manipulated variables for Type I Error study, 2*6*3= 36 conditions and for power analysis 2*6*3*4=144 conditions were generated, and 100 replications were performed for each condition. Before the data was produced, discrimination parameters of the items were defined considering the research design. The multidimensionality of the test was determined according to the discrimination parameters. Accordingly, an item that loads on both dimensions must have two discrimination coefficients. If the item predominantly loads on both dimensions, it is defined as complex; while if it loads dominantly on one dimension and loads little on the other, it is defined as approximately simple, and if it loads dominantly on one dimension and none on the other, it is defined as a simple item. For example, in this study, the first five items of a 10-item test predominantly belong to the first dimension and a small amount to the second dimension while the other five items are arranged in a way that loads predominantly on the second dimension and to a small extent on the first dimension. Thus, a multidimensional test was developed, which predominantly loaded on two different dimensions. While producing the item parameters, ITEM-GENv2 software developed by Ackerman (1994) was used. In this software, parameters are generated by entering only the file name, test length, item angles, the range of the intersection parameter, and the range of the MDISC parameter. Accordingly, items that load on the first dimension make angles with the x-axis that vary between 5° and 20° while items that load on the second dimension make angles that vary between 70° and 88° (Ackerman et al., 2003).

MDISC is the discrimination parameter of multidimensional Item Response Theory (IRT) and corresponds to the item discrimination in unidimensional IRT. Since there is more than one dimension at this point, there is a distinctiveness for each dimension. Item discrimination (MDISC) is represented by a vector (α1, α2, α3… αk). The vector length is expressed as:

$$MDISC = \sqrt{\sum_{n=1}^{k} \alpha^2{}_{ik}} \quad (1)$$

The vector length terms as the common item discrimination (Göçer Şahin, 2016). It could be argued that as the length increases, the discrimination of the item also increases. The α_{ik} in the formula above represents the distinctiveness values of each dimension. The MDISC value here can also be interpreted as distinctiveness in unidimensional IRT.

In addition to the vector length, it is useful to know the vector direction and its distance from the origin. The vector direction is expressed with:

$$\alpha_i = \arccos\left(\frac{\alpha_{i1}}{MDISC}\right) \quad (2)$$

The α_i is the angle that the item vector makes with the θ_1 axis. Thus, an angle of 45° means that the item measures both abilities well. If the angle is greater than 45°, it means that the second dimension is measured better than the first dimension. However, if it is less than 45°, it means that this item primarily measures θ_1 ability, meaning, the first dimension is measured better than the second dimension (Göçer Şahin, 2016; Sünbül, 2011).

In unidimensional IRT, the D parameter is the *b* parameter's equivalent in Multidimensional Item Response Theory (MIRT) and that expresses the distance of the item vector from the starting point and gives information about the item difficulty (Reckase, 2009). This parameter is calculated as:

$$D = \frac{-d_i}{MDISC} \quad (3)$$

The d_i in the formula is described as an intercept term. A negative sign of the item is interpreted as being easy while a positive sign is interpreted as being difficult.

In this study, the range of the MDISC parameter for multidimensional items was entered as 0.8 and 1.8. The study of Ackerman (1994) was taken into account in determining this range. In the condition that the number of simple items is 10 and the structure is bidimensional, the structure of the item, parameters, and the angles of the items with the axes are presented in Table 2 as an example.

Table 2. *Item parameters in data generation.*

Dimensions	*Items*	a_{j1}	a_{j2}	*b*	*MDISC*	*D*	*Angle*
1	1	1.265	.111	-.579	1.27	.46	5.00
	2	1.074	.126	.422	1.08	-.39	6.67
	3	1.671	.245	-.109	1.69	.06	8.33
	4	1.312	.231	-.533	1.33	.40	10.00
	5	.980	.202	-.233	1.00	.23	11.67
	6	.937	.222	-.123	.96	.13	13.33
	7	.903	.242	-.726	.93	.78	15.00
	8	1.164	.349	.415	1.22	-.34	16.67
	9	1.076	.356	.074	1.13	-.07	18.33
	10	.765	.278	-.147	.81	.18	20.00
2	11	.434	1.194	-.579	1.27	.46	70.00
	12	.334	1.029	.422	1.08	-.39	72.00
	13	.465	1.623	-.109	1.69	.06	74.00
	14	.322	1.293	-.533	1.33	.40	76.00
	15	.208	.979	-.233	1.00	.23	78.00
	16	.167	.948	-.123	.96	.13	80.00
	17	.130	.926	-.726	.93	.78	82.00
	18	.127	1.209	.415	1.22	-.34	84.00
	19	.079	1.130	.074	1.13	-.07	86.00
	20	.028	.813	-.147	.81	.18	88.00

2.2. Data Analysis

Both parametric and non-parametric methods were used to compare the performances of various methods in the assessment of unidimensionality. In the scope of this study, the DIMTEST T statistic and Dimensionality DETECT IDN index were used among non-parametric methods. Among parametric methods, Nonlinear Factor Analysis (NOHARM) method was used. The data were analyzed in the following steps:

In the first stage, unidimensional and multidimensional data were generated respectively for testing Type I Errors and power rates. In addition to Stout et al. (1996), Forelich and Habing (2008) studied AT and PT partitioning for the DIMTEST T statistic and (a) it was noted that AT items should be homogeneous in terms of dimensionality, meaning, in terms of geometric representation the angle at which the AT items are located should be relatively narrow. (b) ΘAT and ΘPT should be as different as possible, in other words, in terms of geometric representation the angles between ΘAT and ΘPT should be as large as possible. (c) There must be at least four items in AT while the PT must have at least half of the items in the test. In this study, for the DIMTEST T statistic, AT and PT items were fixed for all conditions, with half of the items in

the AT subtest and the other half in the PT subtest. In the cases where the DIMTEST T statistic was greater than the critical value of 1.96, the Ho hypothesis was rejected.

Dimensionality DETECT IDN index and Nonlinear Factor Analysis methods were used in their default options. Dimensionality DETECT IDN index value of 1 or higher indicates high multidimensionality, while a value between 0.4 and 1 indicates moderate multidimensionality, and a value between 0.2 and 0.4 indicates unidimensionality. In a simulation study by Kim (1994) it was noted that if the Dimensionality DETECT IDN index was less than 0.10, the data could be considered unidimensional. In the same study, it was noted that a value between 0.10 and 0.50 would indicate multidimensionality which was a low probability, a value between 0.51 and 1 would indicate moderate multidimensionality, and a value over 1 would indicate strong multidimensionality (Ackerman & Walker, 2003). 100 replications were performed for all analyses. For each condition of the DIMTEST T statistic and the Dimensionality DETECT IDN index, 4 different result tables were obtained including the reliability coefficients, theta values, the DIMTEST T statistic and the Dimensionality DETECT IDN index. T statistic and p-significance values were reported for the DIMTEST T statistic.

Among parametric methods, nonlinear factor analysis (NOHARM) was applied, and reliability coefficients, theta values and NOHARM result tables were obtained. Two indices, Tanaka Goodness of Fit Index (TIGF) and RMSR, were used to interpret the outputs of the NOHARM program. A TIGF value of ≥ 0.95 and an RMSR value of ≤ 0.05 were evidence of a good fit of the model (Hooper et al., 2008; Hu & Bentler, 1999). In the final step, unidimensionality rejection rates for all outcomes were reported for each condition.

3. FINDINGS

In this section the rates of rejection of unidimensionality as a result of the effect of all conditions and the joint effect of the interaction of the conditions are presented. According to the results of DIMTEST T statistics in Table 3, it was considered that the test length was more inconsistent in making the decision of unidimensionality when the test length was 10 items than when the test length was 20 and 30 items. In addition, in the cases where the test length was 20 and 30, it was considered that it gave more consistent results regardless of the sample size. According to the results of DIMTEST T statistics, regardless of the sample size, as the length of the test increased in unidimensional data, the rate of rejection of unidimensionality generally decreased, in other words, the rate of Type I Error decreased. Another remarkable point in the results of DIMTEST T statistics was that as the sample size increased, the test length produced accurate results for unidimensional data with standard normal distribution, especially in the cases where the test length was 20 and 30 items. It gave more accurate results, especially with a sample size of 300 and above. It could be argued that this finding supports the studies of Finch and Habing (2007) and Finch and Monahan (2008).

When the DETECT IDN index results were examined, the Type I Error rate generally increased as the sample size decreased for the data showing standard normal distribution. Especially when the sample size was 200, 300 and 500, it was noted that the rate of Type I Error was high. However, it could be argued that it gave more inconsistent results when the length of the test was 10 items. In the study conducted by Roussos and Özbek (2006), it was stated that the DETECT IDN index exhibited statistical bias, especially when the test length was 10 or less and the data was unidimensional. Accordingly, the researchers recommended against using DETECT for test lengths of less than 20 items. Although this study coincided with the study of Roussos and Özbek (2006), an important finding was that the sample size should be increased in order to use the DETECT method.

Table 3. *DIMTEST T Statistic, Dimensionality DETECT IDN Index, and Type I Error Rates for Nonlinear Factor Analysis in the data showing normal distribution according to various sample sizes and different numbers of items.*

		DIMTEST T Statistic	DETECT *IDN INDEC*	RMSR	TIGF
Sample Size	*Number of Items*	Rejection rate	Rejection rate	Rejection rate	Rejection rate
200	*10*	0.04	0.71	0.00	0.00
	20	0.00	0.62	0.00	0.00
	30	0.03	0.53	0.00	0.00
300	*10*	0.06	0.64	0.00	0.00
	20	0.00	0.57	0.00	0.00
	30	0.00	0.38	0.00	0.00
500	*10*	0.02	0.59	0.00	0.00
	20	0.00	0.50	0.00	0.00
	30	0.00	0.38	0.00	0.00
1000	*10*	0.05	0.47	0.00	0.00
	20	0.00	0.43	0.00	0.00
	30	0.00	0.31	0.00	0.00
2000	*10*	0.18	0.39	0.00	0.00
	20	0.00	0.30	0.00	0.00
	30	0.01	0.20	0.00	0.00
3000	*10*	0.10	0.38	0.00	0.00
	20	0.00	0.28	0.00	0.00
	30	0.01	0.19	0.00	0.00

Note. N (0,1): Standard Normal Distribution, number of replications: 100, software used for Dimensionality T Statistic: DIMTEST, software used for DETECT IDN index: DETECT, software used for Nonlinear Factor Analysis and Achieved Indexes: NOHARM- RMSR and TIGF

When the RMSR and Tanaka Goodness of Fit Indices were obtained as a result of nonlinear factor analysis that is one of the parametric dimensionality determination methods examined, the Tanaka Goodness of Fit Index (TIGF) value was ≥ 0.95 for unidimensional data with standard normal distribution, regardless of the sample size and the length of the test. However, the RMSR value of ≤ 0.05 in all results proved that the fitness of the model was well. This finding seems to overlap with the study findings of Seo and Sünbül (2012). However, the study by Gessaroli and De Champlain (1996) also showed consistency with conditions where the test length was 15, 30, and 45 items. The DIMTEST T statistic, DETECT IDN index, and Type I Error rates for nonlinear factor analysis in the condition that the test scores were skewed, the sample size was 200, 300, 500, 1000, 2000, and 3000 and the test length was 10, 20 and 30 items are summarized in Table 4.

Table 4. *DIMTEST T Statistic, Dimensionality DETECT IDN Index, and Type I Error Rate for Nonlinear Factor Analysis in skewed data for various sample sizes and number of items.*

		DIMTEST T STATISTIC	DETECT *IDN INDEX*	NOHARM RMSR	NOHARM TIGF
Sample Size	*Number of Items*	Rejection rate	Rejection rate	Rejection rate	Rejection rate
200	*10*	0.04	0.40	0.00	0.00
	20	0.00	0.30	0.00	0.00
	30	0.00	0.31	0.00	0.00
300	*10*	0.05	0.39	0.00	0.00
	20	0.03	0.21	0.00	0.00
	30	0.00	0.29	0.00	0.00
500	*10*	0.04	0.29	0.00	0.00
	20	0.02	0.16	0.00	0.00
	30	0.01	0.18	0.00	0.00
1000	*10*	0.09	0.27	0.00	0.00
	20	0.00	0.03	0.00	0.00
	30	0.02	0.08	0.00	0.00
2000	*10*	0.16	0.17	0.00	0.00
	20	0.01	0.01	0.00	0.00
	30	0.00	0.00	0.00	0.00
3000	*10*	0.18	0.06	0.00	0.00
	20	0.01	0.00	0.00	0.00
	30	0.01	0.00	0.00	0.00

Note. (1.75, 3.75) Skewed Distribution, number of replications:100, software used for Dimensionality T Statistic: DIMTEST, software used for DETECT IDN index: DETECT, Software used for Nonlinear Factor Analysis and Indexes: NOHARM-RMSR and TIGF

According to Table 4, the Type I Error rate was particularly higher in small samples and in the cases when test length was short, and the distribution was skewed. Although it was noted that the DIMTEST T statistic gave more accurate results than DETECT IDN index, it was found that the error rate was higher in the DIMTEST T statistic results when the test length was 10 items compared to other test lengths. However, in all conditions where the test length was 20 and 30 items, it was noted that the DIMTEST T statistic gave very accurate results. When the nonlinear factor analysis (NOHARM) results were examined, it showed a rejection rate of 0.00 for unidimensional data with skewed distribution, regardless of the sample size and the test length. The findings of the third group of the study were in conditions where the data had standard normal distribution, the sample sizes were 200, 300, 500, 1000, 2000 and 3000, and the test length was 10, 20, and 30 items and there was an interdimensional correlation with 0.25, 0.50, 0.75, and 0.90. The power rates of the test for DIMTEST T statistic, the Dimensionality DETECT IDN index, and the Nonlinear Factor Analysis (NOHARM) results are summarized in Table 5:

Table 5. *Power rates for DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis in data with standard normal distribution according to various sample sizes, different numbers of items, and different interdimensional correlations.*

N~(0,1)		Interdimensional Correlation															
		0.25				0.50				0.75				0.90			
		Rejection Rate				Rejection Rate				Rejection Rate				Rejection Rate			
Sample Size	*Number of Items*	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF
200	*10*	0.91	0.86	1.00	1.00	0.64	0.92	1.00	1.00	0.37	0.92	1.00	1.00	0.14	0.96	1.00	1.00
	20	1.00	1.00	1.00	1.00	0.87	1.00	1.00	1.00	0.29	1.00	1.00	1.00	0.08	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.52	1.00	1.00	1.00	0.13	1.00	1.00	1.00
300	*10*	0.99	0.74	1.00	1.00	0.95	0.85	1.00	1.00	0.55	0.91	1.00	1.00	0.26	0.97	1.00	1.00
	20	0.99	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.66	1.00	1.00	1.00	0.23	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.86	1.00	1.00	1.00	0.27	1.00	1.00	1.00
500	*10*	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.69	0.99	1.00	1.00	0.30	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.73	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.47	1.00	1.00	1.00
1000	*10*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.56	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.83	1.00	1.00	1.00
2000	*10*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
3000	*10*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	98	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. N (0,1): Standard Normal Distribution, number of replications:100, software used for Dimensionality T Statistic: DIMTEST, software used for DETECT IDN index: DETECT, Software used for Nonlinear Factor Analysis and Achieved Indexes: NOHARM- RMSR and TIGF

According to the results of the DIMTEST T statistics, in the data showing standard normal distribution, in the case of an interdimensional correlation of 0.25 and with a sample size of 500 and above, no matter what the length of the test was, the unidimensionality in bidimensional data showed standard normal distribution for all conditions while the rejection rate was found to be 1.00. The rejection rate for unidimensionality was found to be 1.00 in all conditions, except for the condition where the interdimensional correlation was 0.50, the sample size was 500 and the test length was 10 items. In addition, for the two conditions (200 and 300) where

the sample size was less than 500, the rejection rate of unidimensionality was lower than in the cases with larger sample sizes. As a result, it could be argued that the DIMTEST T statistic gave more accurate results in conditions with large samples. This finding is consistent with the studies of Finch and Habing (2007), Finch and Monahan (2008), and Özbek Baştuğ (2012). Especially in the cases where the sample size was less than 300, the error rate of DIMTEST T statistics increased significantly. According to the results of DIMTEST T statistics, as the interdimensional correlation value increased, the unidimensionality rejection rate in bidimensional data decreased. In other words, the power of the test decreased. In the cases where the interdimensional correlation was low, the rejection rate of unidimensionality was 1.00 for the DIMTEST T statistic regardless of the sample size and the test length. In other words, the data was accepted to be bidimensional and the power of the test was high. It was noted that the DIMTEST T statistic was significantly affected by the interdimensional correlation for the multidimensionality decision. However, in the cases when the sample size was 3000 and the test length was 30, regardless of the correlation value between dimensions, a rejection rate of 1.00 was achieved for unidimensionality. In other words, an excellent decision was made for multidimensionality. In the study conducted by Zhang (2008), it was stated that in the condition of low interdimensional correlation, short tests produced better results than long tests. However, in this study, when the interdimensional correlation was very low, the results of DIMTEST T statistics gave an excellent performance in terms of test power as the test length increased. Although the result of this study was inconsistent with the study of Zhang (2008), it seemed to overlap with the studies by Alexandra et al. (2004), Seo and Sünbül (2012) and Özbek Baştuğ (2012).

When the results of the dimensionality DETECT IDN index for the power of the test were examined, in the case of bidimensional data with standard normal distribution, with a sample size of 500 and above, the correlation value between dimensions and the test length displayed a rejection rate of 1.00 for all conditions except one. It could be argued that the Dimensionality DETECT IDN statistic worked well in rejecting unidimensionality and accepting bidimensionality in cases with bidimensional data where the sample size was 500 and above. This finding was consistent with the findings of the study by Svetina (2011) and the studies of Roussos and Özbek (2006). In the data with standard normal distribution, when the RMSR and Tanaka Goodness of Fit Index values were examined following nonlinear factor analysis (NOHARM) as a parametric method for test power, it was observed that for bidimensional data with standard normal distribution, interdimensional correlation displayed a rejection rate of 1.00 for unidimensionality, regardless of sample size and test length. In other words, the null hypothesis that the test was unidimensional in all circumstances was correctly rejected. When the relevant literature was reviewed, it was stated in the study conducted by Kaya and Kelecioğlu (2016) that the results of nonlinear factor analysis were more consistent in determining multidimensionality in samples of 50 or more. Contrary to this study, studies by Özbek Baştuğ (2012) and Seo and Sünbül (2012) found that nonlinear factor analysis (NOHARM) was not a powerful statistical method for determining multidimensionality. However, Svetina (2011) stated that statistics based on nonlinear factor analysis (NOHARM) results in determining dimensionality in data suitable for non-compensatory multidimensional IRT models showed a stronger performance compared to Dimensionality DETECT IDN index.

As a result, it was noted that the dimensionality DETECT IDN index and nonlinear factor analysis (NOHARM) results gave more accurate decisions than the DIMTEST T statistic under all conditions in the data with standard normal distribution. It could be argued that the DIMTEST T statistic gave more accurate decisions in conditions where the interdimensional correlation was low, and the sample size was large. In addition, it could be argued that the DIMTEST T statistic worked better in samples of 2000 and above in the cases where the interdimensional correlation was high.

The findings for the 4th group of the study are presented in Table 6. Accordingly, the DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis (NOHARM) results were compared in terms of test power ratios in the data with skewed distribution, where the sample size was 200, 300, 500, 1000, 2000, and 3000, the test length was 10, 20, and 30 items, and the degree of interdimensional correlation was 0.25, 0.50, 0.75, and 0.90.

Table 6. *Power rates for DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis in data with skewed distribution according to various sample sizes, different numbers of items, and different interdimensional correlation values.*

		Interdimensional Correlation															
		0.25				0.50				0.75				0.90			
Sample Size	*Number of Items*	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT *IDN INDEX*	RMSR	TIGF
200	*10*	0.84	0.76	1.00	1.00	0.69	0.79	1.00	1.00	0.24	0.86	1.00	1.00	0.14	0.85	1.00	1.00
	20	0.96	1.00	1.00	1.00	0.89	1.00	1.00	1.00	0.36	0.99	1.00	1.00	0.08	0.99	1.00	1.00
	30	0.90	1.00	1.00	1.00	0.82	1.00	1.00	1.00	0.61	1.00	1.00	1.00	0.13	1.00	1.00	1.00
300	*10*	0.98	0.67	1.00	1.00	0.91	0.71	1.00	1.00	0.61	0.67	1.00	1.00	0.26	0.77	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.23	0.99	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	0.27	1.00	1.00	1.00
500	*10*	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.77	1.00	1.00	1.00	0.73	0.99	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	0.68	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.40	1.00	1.00	1.00
1000	*10*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.68	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.56	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.82	1.00	1.00	1.00
2000	*10*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00
3000	*10*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00

According to Table 6, in the data with skewed distribution when the sample size increased and the number of items in the test increased and the power ratios for DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis were analyzed according to different interdimensional correlation values, it was noted that all conditions in which nonlinear factor analysis and Dimensionality DETECT IDN index were used gave more accurate decisions than the DIMTEST T statistics. However, according to the DIMTEST T statistic, it could be argued that more accurate decisions were made in conditions when interdimensional correlation was low. In addition, in the cases when the interdimensional correlation was high, it was noted that the DIMTEST T statistic worked better in samples of 1000 and above.

In the conditions where the sample size was 200 and 300 and the test length was 10 items, it was noted that the rate of correct decision-making decreased in the results of the Dimensionality DETECT IDN index and the DIMTEST T statistics, regardless of the interdimensional correlation. Although the correct decision rate of DETECT IDN index and the DIMTEST T statistic increased as the test length increased, it could be argued that the correct decision rate of the DIMTEST T statistic decreased as the sample size decreased. It could be argued that nonlinear factor analysis worked better than the Dimensionality DETECT index and DIMTEST T statistics in the process of determining dimensionality with skewed data.

4. DISCUSSION and CONCLUSION

When the DIMTEST T statistic, Dimensionality DETECT IDN index, and Type I Error rates for Nonlinear Factor Analysis were examined in data with standard normal distribution, according to various sample sizes and different item numbers, the Nonlinear Factor Analysis (NOHARM) results were the most consistent in making the unidimensionality decision. In addition, although results of the DIMTEST T statistics were argued to be more consistent, it was thought that the use of DIMTEST T statistics in determining dimensionality in short tests would not be appropriate. In addition, it could be argued that the DETECT IDN index would be more appropriate to use with large samples and large test lengths. The DETECT IDN index should not be used in the dimensionality determination process, especially in short tests. When the DIMTEST T statistic, Dimensionality DETECT IDN index, and Nonlinear Factor Analysis (NOHARM) Type I Error rates were examined according to various sample sizes and different numbers of items with the data showing skewed distribution, it was observed that the results of DETECT IDN index were more consistent with the data showing skewed distribution compared to the data showing standard normal distribution. The results of DIMTEST T statistics and Nonlinear Factor Analysis were found to be more accurate in making the unidimensionality decision.

When the power rates for the DIMTEST T Statistics, dimensionality DETECT IDN index and Nonlinear Factor Analysis were examined according to various sample sizes, different numbers of items and different interdimensional correlation values in the data with standard normal distribution, it could be argued that the DIMTEST T statistic gave more accurate results in conditions with large samples. Especially in the cases when the sample size was less than 300, the error rate of DIMTEST T statistics increased significantly. At the same time, it could be argued that the DIMTEST T statistic was affected by the interdimensional correlation for the multidimensionality decision. In data with standard normal distribution, the results of the dimensionality DETECT IDN index and nonlinear factor analysis (NOHARM) seemed to make more accurate decisions than the DIMTEST T statistic under all conditions. DIMTEST T statistic, on the other hand, was found to make more accurate decisions in conditions with low interdimensional correlation and high sample sizes.

It could be argued that dimensionality determination methods gave less consistent results when the test length was less than 10 items with skewed distribution. On the other hand, although it was seen that the results of DETECT IDN index and Nonlinear factor analysis had higher inner consistency, it could be argued that the DIMTEST T statistic gave the right decisions in determining dimensionality when the sample size was 1000 and above.

As in every study, this study also had some limitations. The conditions discussed in this study were limited to sample size (200, 300, 500, 1000, 2000, and 3000), interdimensional correlation (0.25, 0.50, 0.75, 0.90), test length (10, 20, 30 items), and different ability distributions (standard normal distribution and skewed distribution). A similar study could be repeated with smaller samples and conditions with a larger test length. In addition, the research could be repeated by adding other variables. Based on the results of the DIMTEST T statistic used in

this study together with DETECT IDN index and nonlinear factor analysis and considering item pools and large samples of the large-scale tests used in the exams administered by the Student Selection and Placement Center (ÖSYM) or the Ministry of National Education (MEB), use of nonlinear factor analysis, the DIMTEST T statistic, and DETECT IDN index were found suitable to determine their dimensionality. In addition, nonlinear factor analysis seems to be a more accurate decision, especially instead of DETECT IDN index and the DIMTEST T statistic, in determining the dimensionality of short exams applied in the school environment.

In this study, 2PL and compensatory models were used. In future studies, together with 3PL models, the results can be examined using non-compensatory models, especially for tests containing items where one dimension does not compensate for the other dimension. In this study, test cases that were scored 1-0 were created. Considering the scale development and scale adaptation studies in education and psychology in future studies, the effectiveness of the same methods can be investigated in tests with multiple scores.

A similar study can be conducted by increasing the number of dimensions. The efficiency of the methods can also be tested on real data in the same study. The structure of the test discussed in this study is fixed and the test is semi-mixed. A similar study can be conducted with a different structure by varying the number of simple or complex items and different test structures can be used to test the effect of the test structure. Different item parameter sets can affect the performance of methods. Thus, in order to make the findings more generalizable, it could be useful to compare the present results with results based on a different set of item parameters. Considering the answers not given in the test items used in the exams held in our country, the efficiency of the methods can be tested by manipulating the amount of missing data in another study. While creating the skewed distribution in this study, skewness and kurtosis values (1.75, 3.75) in Fleisman's (1978) study were taken into account. Data set could be created considering the different deviations from the standard normal distribution, and the Type I Error and power study could be assessed for the dimensionality determination process.

In this study, Nonlinear Factor Analysis (NOHARM) from among parametric dimensionality determination methods and the DIMTEST T statistic from among non-parametric methods and Dimensionality DETECT IDN Index were used. In a different study, performances of other parametric and non-parametric methods in dimensionality determination can be tested. Among the parametric and non-parametric methods selected for the scope of this study, indices such as RMSR, Tanaka Goodness of Fit Index, and DETECT IDN index were used. In a different study, the Type I Error and power study can be assessed using other indices such as the approximate chi-square ($\chi^2_{G/D}$) statistic index obtained using the same methods. One of the important results of this study is that authors should consider the strengths and weaknesses of the methods in terms of the characteristics of the data while deciding or choosing the methods for determining dimensionality. Considering the difficulties in data collection processes, especially in the field of social sciences in our country, studies should be conducted using recommended methods in order not to reach inconsistent results due to the effect of sample size. Finally, for authors that would like to conduct a determination of dimensionality studies in the cases where research has not yet proven the superiority of one method over another, the application of multidimensionality methods may be useful if authors would like to have a comprehensive understanding of structure and dimensionality of the data before moving on to the scores obtained from the tests.

Acknowledgments

This paper was produced from the part of the first author's doctoral dissertation prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Gul Guler: Investigation, Software, Methodology, Formal Analysis, Visualization, Resources, and Writing the original draft. **Rahime Nukhet Cikrikci:** Software, Methodology, Supervision, and Validation.

Orcid

Gul Guler https://orcid.org/0000-0001-8626-4901
Rahime Nukhet Cikrikci https://orcid.org/0000-0001-8853-4733

REFERENCES

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278. https://doi.org/10.1207/s15324818ame0704_1

Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Davey, T., Nering M.L., & Thompson, T. (1997). Realistic simulation of item response data. ACT Research Report Series, 97-4. https://files.eric.ed.gov/fulltext/ED414297.pdf

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

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARMbased statistics for testing unidimensionality. *Applied Psychological Measurement, 31*, 292-307. https://doi.org/10.1177/0146621606294490

Fleisman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532. https://doi.org/10.1007/BF02293811

Froelich, A.G., & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement, 32*, 138-155. https://doi.org/10.1177/0146621607300421

Gessaroli, M.E., & De Champlain, A.F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, 33, 157-179. https://doi.org/10.1111/j.1745-3984.1996.tb00487.x

Göçer Şahin, S. (2016). *Yarı karışık yapılı çok boyutlu yapıların tek boyutlu olarak ele alınması durumunda kestirilen parametrelerin incelenmesi [Examining parameter estimation when treating semi-mixed multidimensional constructs as unidimensional]* [Unpublished doctoral dissertation]. Hacettepe University.

Hattie, J. (1985). Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9(8),* 139 – 145. http://dx.doi.org/10.1177/014662168500900204

Hattie, J., Krakowski, K., Rogers, H.J., & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement, 20,* 1-14. https://doi.org/10.1177%2F014662169602000101

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods 6*(1), 53-60. https://doi.org/10.21427/D7CF7R

Hu, L-T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: *Conventional criteria versus new alternatives. Structural Equation Modeling*, 6, 1-55. https://doi.org/10.1080/10705519909540118

Kaya, K.Ö., & Kelecioğlu, H. (2016). The effect of sample size on parametric and nonparametric factor analytical methods. *Educational Sciences: Theory & Practice. 16*(1), 153-171. http://dx.doi.org/10.12738/estp.2016.1.0220

Kim, H.R. (1994). New techniques for dimensionality assessment of standardized test data. [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign, Department of Statistics.

Ledasma, R.D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation*, *12* (2). https://doi.org/10.7275/wjnc-nm63

Mroch, A.A., & Bolt, D.M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education, 19* (1), 67-91. https://doi.org/10.1207/s15324818ame1901_4

Nandakumar, R., & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*(1), 41-68. https://psycnet.apa.org/doi/10.2307/1165182

Özbek Baştuğ, Ö.Y. (2012). Assessment of Dimensionality in Social Science Subtest. *Educational Sciences: Theory & Practice. 12(1), Winter:* 382-385.

Reichenberg, R.E. (2013). *A comparison of DIMTEST and generalized dimensionality discrepancy approaches to assessing dimensionality in item response theory* [M.S. dissertation, Arizona State University, Arizona]. https://doi.org/10.3102%2F10769986018001041

Reckase, M.D. (2009). *Multidimensional item response theory.* Springer Dordrecht Heidelberg.

Roussos, L.A., & Özbek, O.Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement, 43*, 215-243. https://doi.org/10.1111/j.1745-3984.2006.00014.x

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617. https://doi.org/10.1007/BF02294821

Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L., & Zhang J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 19*, 331-354. https://doi.org/10.1177%2F014662169602000403

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi [Examining item parameter invariance for several dimensionality types by using classical test theory, unidimensional item response theory and multidimensional item response theory]* [Unpublished doctoral dissertation]. Mersin University.

Sünbül, Ö., & Seo, M. (2012). *Performance of test statistics for verifying unidimensionality,* [Conference presentation abstract]. 2012 Annual Meeting, April 12-16, Vancouver, British Columbia, CANADA

Svetina, D. (2011). Assessing dimensionality in complex data structures: A performance comparison of DETECT and NOHARM procedures [Unpublished doctoral dissertation]. Arizona State University.

Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment, 19(1),* 35-57. https://doi.org/10.1080/10627197.2014.869450

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159-203. https://doi.org/10.1177/0146621603027003001

Touron, J., Lizasoain, L., & Joaristi, L. (2012). Assessing the unidimensionality of the School and College Ability Test (SCAT, Spanish version) using non-parametric methods based on item response theory. *High Ability Studies.* 23(2), 183-202. https://doi.org/10.1080/13598139.2012.735401

Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimension. *The Journal of Experimental Education, 77* (2), 147-166. https://doi.org/10.3200/JEXE.77.2.147-166