International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 263–273

DOI: 10.21449/ijate.377138

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

Effects of Various Simulation Conditions on Latent-Trait Estimates: A Simulation Study

Hakan Kogar [1*]

[1] Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Turkey

Abstract: The aim of this simulation study, determine the relationship between true latent scores and estimated latent scores by including various control variables and different statistical models. The study also aimed to compare the statistical models and determine the effects of different distribution types, response formats and sample sizes on latent score estimations. 108 different data bases, comprised of three different distribution types (positively skewed, normal, negatively skewed), three response formats (three-, five- and seven-level likert) and four different sample sizes (100, 250, 500, 1000) were used in the present study. Results show that, distribution types and response formats, in almost all simulations, have significant effect on determination coefficients. When the general performance of the models are evaluated, it can be said that MR and GRM display a better performance than the other models. Particularly in situations when the distribution is either negatively or positively skewed and when the sample size is small, these models display a rather good performance.

ARTICLE HISTORY

Received: 10 January 2018
Revised: 03 March 2018
Accepted: 16 March 2018

KEYWORDS

Item response theory,
Classical test theory,
Factor analysis,
Latent trait scores,
Data simulation

1. INTRODUCTION

In the Classical Test Theory (CTT), known to be the first theory developed to measure latent traits, the fundamental concept is the true score. The true score is defined as the expected value of the observed scores. The expected value expressed in this definition can be obtained by means of an infinite number of repetitions of the independent observations (Lord & Novick, 1968). In other words, if a psychological test is to be administered, the test taker's true score can be obtained by administering the test to the person an infinite number of times. According to this theory, the mathematical representation of which is rather simple, the observed score is obtained by adding the true score and the random error (Mellenberg, 1996). The latent score in CTT refers to the observed scores obtained by adding the item scores (Lord & Novick, 1968).

Item Response Theory (IRT), known to be a modern test theory, was developed based on the argument that it is not realistic to make infinite observations and that repeated measurements are not statistically independent of each other. IRT and CTT are different in

CONTACT: Hakan Kogar hkogar@gmail.com Akdeniz University Education Faculty Measurement and Assessment Department Antalya Turkey

ISSN-e: 2148-7456 /© IJATE 2018

terms of their theoretical basics and statistical formulations (Borsboom & Mellenbergh, 2002). When both are compared, it is believed that IRT is superior as psychometric traits can be obtained independent of the sample and to which test or item an ability or trait belongs to can be determined from the participants' responses (Crocker & Algina, 1986). IRT models seek to determine the latent traits based on their item stimulators (such as item difficulty and estimate of parameters) and the interaction of the ability. In these models, instead of the total score, the patterns in the responses are focused on. IRT, which is widely used in the fields of education and psychology, has various latent trait models which can be applied to dichotomous or polytomous datasets (Brzezińska, 2016).

While IRT models make use of all the information in the response patterns in order to obtain all the item parameters, factor analysis (FA) techniques estimate the relationships between items and latent traits by means of correlation matrices (Cyr & Davies, 2005). Principal component analysis (PCA), which is considered as the basic method of factor analysis, is a dimension reduction method. It seeks to derive a small number of independent principal components from a larger number of correlated variables (Saporta & Niang, 2009). While latent variables can directly be measured in PCA, in factory analysis, data reduction can only be used for traits that cannot be directly measured (e.g. intelligence, anxiety). A theoretical definition is needed for these traits that cannot be directly measured (Bartholomew, Knott, & Moustaki, 2011). Researchers who seek to determine how many factors have an effect on a variable and which factors have a combined effect utilize exploratory factor analysis (EFA), which is based on an exploratory technique (DeCoster, 1998). When the relationship between the observed and latent variables is revealed, confirmatory factor analysis (CFA) is used. CFA is a measurement model that seeks to estimate the population covariance matrix of the theoretical model based on the observed covariance matrix (Raykoy & Marcoulides, 2000, 95).

Not many studies are encountered in the related literature which comparisons are made between the different parameter estimation methods on these techniques, namely CTT, IRT, and FA (Dumenci & Achenbach, 2008; Hauck Filho, Machado, & Damásio, 2014). In one study, conducted by Dumenci and Achenbach (2008), six statistical models that could estimate different latent traits were compared: CTT, PCA, CFA using maximum likelihood estimation, CFA using weighted least squares, graded response model (GRM) and partial credit model (PCM). CTT, PCA and CFA using the maximum likelihood estimation method yielded similar findings. Likewise, similar findings were observed among the PCA, GRM and CFA using weighted least squares models. In each group of methods, the estimations of the linear relationships (r^2) were found to be close to 1.00. As real data were used in the study, the lack of control variables made it difficult for the models to be compared. In another study, conducted by Hauck Filho et al. (2014), seven different statistical models that could estimate latent traits were compared: CTT, PCA, EFA using Maximum Likelihood, EFA with Minimum Rank, RSM, GRM and CFA with weighted least squares. This comparative study was performed with a total of 15 different simulative datasets comprised of three different item difficulty distributions and five different sample sizes. In each dataset, based on 10 items, true scores of latent traits were obtained. The comparison between the true scores and the estimated trait scores were tested by means of various statistical techniques. It was found that the estimations that were closest to the true scores were those estimations obtained from RSM, GRM and CFA using weighted least squares. These three models are ones that are least affected by inconsistencies among the items and sample distributions. However, the findings of these three models were not found to be statistically significant.

The present simulation study, which took into consideration previous studies, aimed to determine the relationship between true latent scores and estimated latent scores by including various control variables (distribution types and response formats) and different statistical

models (unweighted least squares and diagonally weighted least squares). The study also aimed to compare the statistical models and determine the effects of different distribution types, response formats and sample sizes on latent score estimations.

2. METHOD

2.1. Procedures of Data Simulations

Based on three different item difficulty distributions (which is defined below), 108 different data bases, comprised of three different distribution types (positively skewed, normal, negatively skewed), three response formats (three-, five- and seven-level likert) and four different sample sizes (100, 250, 500, 1000) were used in the present study. In these data bases, the discrimination parameter (parameter a) was kept constant between 0.5 and 2.8 owing to the fact that the distribution of the simulative datasets was similar to that of the true datasets. The item responses were produced via the Generalized Partial Credit Model (GPCM). Ability parameters (theta) were calculated for each database. These values were recorded as true latent scores. Total of 20 items were simulated.

Among the three different item difficulty distributions, the first (Situation-1) aimed to include the individuals who were in the lower 20% of the sample distribution, that is between -3.00 and -0.84 in terms of the item difficulty parameter (parameter b). The second item difficulty distribution (Situation-2) was simulated with a standard normal distribution having a mean of 0 and a standard deviation of 1. The third item difficulty distribution (Situation-3) included the individuals in the top 20% of the sample distribution that is between 0.84 and 3.00 in terms of the item difficulty parameter (parameter b). These values were obtained by means of the z-score table. These values are adapted from Hauck Filho, et al. (2014).

Of the three different distribution types, the first was a negatively skewed distribution. Taking into consideration beta distribution, this distribution was produced with an expected skewness of 0.40 and an expected kurtosis of -0.30. For this purpose, in the beta distribution, value a was 5.7 and value b was 2.9. The normal distribution, which is the second distribution type, was mean of 0 and the standard deviation of 1. Taking into consideration beta distribution, the positively skewed distribution, which was the third distribution type, was produced with an expected skewness of 0.40 and an expected kurtosis of -0.30. For this purpose, in the beta distribution, value a was 2.9 and value b was 5.7. These values are adapted from Hauck Filho, et al. (2014).

The difference in the sample size was determined, considering previous simulation studies (Dawber, Rogers, & Carbonaro, 2009; Hauck Filho, et al., 2014). Even though one of the factors affecting the psychometric traits of measurement instruments is the response formats (Jafari, Bagheri, Ayatollahi, & Soltani, 2012), the same number of response formats was used in almost all simulation studies. However, there are simulation studies that seek to determine the most appropriate response format for psychological measurement instruments. The response formats in the present study were determined by taking into consideration the findings of studies in which the most appropriate number of response categories was stated (Lozano, García-Cueto, & Muñiz, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009). Data simulation was implemented using the WINGEN program (Han, 2007).

2.2. Data Analysis

In the present study, latent trait score estimates were made by means of the different models stated below:

Classical Test Theory (CTT): In congruence with this theory, for every database, the raw scores (total score) were calculated based on a 20-item test.

Principal Component Analysis (PCA): Component scores were obtained by using this method, which produced weighted scores from indicators (items). Regression scoring method was used for estimate. Factor scores were obtained using the Factor 10.5 program.

Minimum Rank Factor Analysis (MR): This parameter estimation method was developed by Ten Berge and Kiers (1991) with the purpose of explaining the common variance at the highest level. By using the Factor 10.5 program and this parameter estimation method, the polychoric correlation matrix (Lorenzo-Seva & Ferrando, 2006) and the factor scores were determined.

Unweighted Least Squares (ULS): With this method, which can independently make parameter estimations based on distribution types (Kline, 2015, p. 159), a confirmatory factory analysis was conducted. The factor values were obtained via LISREL 8.7.

Diagonally Weighted Least Squares (DWLS): DWLS is a CFA model specifically designed for ordinal data. DWLS does not have any distribution assumptions (Li, 2016). The factor values were obtained via LISREL 8.7.

Graded Response Model (GRM): This model, which is a IRT method used in multiple score scales, such as Likert type scales (Samejima, 1968), was used in combination with estimated a posteriori (EAP) and the R 3.4.2 program and the psych (Revelle, 2017) and Itm (Revelle, 2017) packages to estimate ability parameters.

The Pearson correlation coefficients and determination coefficients (r^2) between the obtained latent trait estimates (scores and indices) and the true latent scores were obtained. In addition, in all the simulation conditions, the factorial ANOVA test was run to test the mean differences and the common variance.

3. FINDINGS

The relationship between six different methods used to estimated latent trait scores and true latent scores in a total of 108 different simulative datasets consisting of three different item difficulty distributions, three different distribution types, three different response formats and four different sample sizes, and the findings regarding determination coefficients are presented in Tables 1, 2 and 3.

In Situation-1, there were huge differences between the correlation and determination coefficients obtained from the negative skewed distribution. Particularly in sample size-1 and response format-1 conditions, zero correlation was found between the true score and the latent trait scores that the models yielded. Nor was zero correlation found for sample size-1 and response format-3. It was found that there was a high correlation between latent trait estimates obtained via a negatively skewed distribution in MR and true scores only in sample size-1 and response format-4, while the relationships in the other simulation conditions were close to zero. The estimations of the other five models yielded moderate or high correlation coefficients in the other simulation conditions. CTT produced a correlation coefficients with the highest average. In the normal distribution in Situation-1, the correlation coefficients in all the simulation conditions were moderate or high. The estimations that the MR model yielded had correlation coefficients with the highest average. In the positively skewed distribution in Situation-1, the correlation coefficients obtained in all the simulation conditions were very high ($r>.88$). The estimations that GRM yielded had a correlation coefficients with the highest average.

The correlation coefficients obtained in the simulation condition with a negatively skewed distribution (Situation-2), except for the estimations made for sample size-1 and response format-1 via MR model, were found to be very high ($r>.90$). It was observed that the estimations obtained via the MR model were affected by a negatively skewed distribution, particularly in situations with a small sample size. It was also found that in a simulative

database obtained from a normal distribution, it was the MR model estimations that were mostly affected, but all the models yielded estimations with high correlation coefficients. It was found that in positively skewed distributions, the estimations that the DWLS model yielded were affected by small sample sizes. In Situation-2, the higher the response format and sample size were, the higher the correlations and determination coefficients turned out to be. In Situation-2, the estimations that GRM yielded in all conditions had coefficients of relationship with the highest averages.

Table 1. Correlation and determination coefficients for situation-1

D	RF	S	Situation-1											
			CTT		PCA		MR		ULS		DWLS		GRM	
			R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
D-1	RF -1	S1	.016	.000	.006	.000	-.008	.000	-.022	.000	.055	.003	.015	.000
		S2	.753	.567	.681	.463	.635	.403	.687	.472	.695	.483	.676	.457
		S3	.049	.002	.039	.001	.048	.002	.037	.001	.034	.001	.048	.002
		S4	.696	.484	.660	.435	.701	.492	.654	.428	.658	.433	.642	.413
	RF -2	S1	.720	.519	.642	.413	.009	.000	.642	.413	.642	.413	.458	.209
		S2	.707	.499	.645	.415	.024	.001	.645	.415	.572	.328	.687	.472
		S3	.706	.499	.665	.443	-.037	.001	.654	.428	.621	.386	.637	.406
		S4	.692	.479	.617	.381	.040	.002	.615	.378	.569	.324	.734	.539
	RF -3	S1	.699	.488	.675	.455	-.091	.008	.655	.429	.559	.313	.537	.288
		S2	.634	.401	.590	.348	-.051	.003	.559	.312	.451	.203	.695	.482
		S3	.692	.479	.667	.445	-.116	.013	.652	.425	.651	.424	.762	.580
		S4	.717	.513	.674	.454	.042	.002	.666	.444	.629	.396	.778	.605
Δ			.737	.567	.675	.463	.817	.492	.709	.472	.661	.482	.763	.605
Mean			.590	.411	.547	.354	.100	.077	.537	.345	.511	.309	.556	.371
D-2	RF -1	S1	.849	.721	.827	.684	.818	.669	.826	.682	.823	.678	.881	.849
		S2	.801	.641	.768	.589	.813	.661	.764	.584	.755	.570	.727	.529
		S3	.837	.700	.817	.668	.856	.733	.812	.660	.791	.626	.887	.787
		S4	.834	.695	.819	.670	.885	.783	.815	.665	.811	.658	.902	.813
	RF -2	S1	.766	.586	.753	.568	.837	.701	.747	.558	.652	.425	.815	.766
		S2	.784	.615	.729	.532	.874	.763	.711	.506	.715	.512	.829	.687
		S3	.788	.621	.773	.597	.847	.717	.774	.599	.772	.597	.827	.683
		S4	.776	.603	.745	.555	.866	.749	.750	.562	.748	.559	.864	.746
	RF -3	S1	.816	.666	.814	.662	.898	.807	.803	.644	.813	.661	.844	.816
		S2	.787	.619	.775	.601	.897	.804	.785	.616	.788	.621	.856	.733
		S3	.788	.621	.775	.601	.900	.810	.769	.591	.766	.587	.859	.738
		S4	.778	.606	.773	.597	.883	.779	.765	.585	.766	.587	.887	.788
Δ			.083	.135	.098	.152	.087	.149	.115	.176	.171	.253	.175	.320
Mean			.800	.641	.781	.610	.865	.748	.777	.604	.767	.590	.848	.745
D-3	RF -1	S1	.907	.823	.904	.816	.903	.815	.900	.811	.898	.806	.937	.907
		S2	.914	.836	.913	.834	.920	.846	.915	.837	.914	.835	.946	.894
		S3	.902	.813	.902	.814	.925	.855	.905	.819	.905	.820	.933	.870
		S4	.906	.820	.903	.816	.927	.860	.906	.820	.905	.819	.938	.880
	RF -2	S1	.897	.805	.893	.798	.934	.872	.890	.792	.887	.787	.941	.897
		S2	.936	.876	.934	.872	.955	.911	.934	.871	.933	.871	.962	.926
		S3	.911	.829	.905	.819	.943	.890	.906	.821	.889	.791	.949	.901
		S4	.910	.827	.908	.824	.944	.891	.909	.826	.909	.826	.951	.905
	RF -3	S1	.917	.842	.914	.836	.946	.895	.917	.840	.914	.835	.951	.917
		S2	.910	.829	.909	.826	.947	.897	.911	.830	.913	.834	.958	.917
		S3	.890	.793	.887	.787	.951	.904	.881	.776	.883	.780	.951	.904
		S4	.912	.833	.908	.825	.953	.908	.908	.824	.908	.825	.958	.917
Δ			.046	.083	.047	.085	.052	.096	.053	.095	.050	.091	.029	.056
Mean			.909	.827	.907	.822	.937	.879	.907	.822	.905	.819	.948	.903

D: Distribution type, RF: Response format, S: Sample Size

Table 2. Correlation and determination coefficients for situation-2

D	RF	S	Situation-2											
			CTT		PCA		MR		ULS		DWLS		GRM	
			R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
D-1	RF -1	S1	.937	.877	.926	.857	.779	.606	.934	.873	.935	.874	.950	.902
		S2	.938	.879	.934	.872	.911	.829	.939	.881	.941	.885	.915	.837
		S3	.940	.883	.933	.871	.906	.822	.933	.870	.929	.863	.956	.915
		S4	.938	.880	.937	.878	.930	.865	.940	.883	.939	.882	.945	.894
	RF -2	S1	.961	.924	.962	.925	.964	.929	.963	.926	.959	.919	.979	.959
		S2	.949	.901	.948	.898	.940	.884	.942	.887	.944	.891	.970	.941
		S3	.956	.913	.955	.911	.956	.914	.953	.909	.952	.907	.973	.947
		S4	.960	.922	.959	.920	.964	.929	.962	.926	.963	.927	.972	.944
	RF -3	S1	.965	.931	.963	.928	.975	.951	.960	.921	.954	.910	.985	.970
		S2	.954	.910	.948	.900	.972	.944	.951	.904	.951	.904	.970	.940
		S3	.963	.928	.963	.927	.969	.939	.962	.926	.963	.927	.977	.954
		S4	.962	.926	.962	.926	.972	.944	.962	.926	.961	.924	.980	.961
Δ			.028	.054	.037	.071	.196	.345	.030	.056	.034	.064	.070	.133
Mean			.952	.906	.949	.901	.937	.880	.950	.903	.949	.901	.964	.930
D-2	RF -1	S1	.942	.887	.944	.892	.883	.779	.943	.890	.943	.889	.946	.895
		S2	.959	.920	.962	.925	.947	.897	.962	.926	.961	.924	.973	.947
		S3	.946	.895	.949	.900	.919	.845	.946	.896	.947	.896	.957	.916
		S4	.947	.896	.950	.902	.954	.909	.950	.903	.950	.902	.963	.927
	RF -2	S1	.971	.942	.971	.942	.973	.947	.968	.938	.967	.935	.977	.955
		S2	.963	.927	.965	.932	.980	.961	.963	.927	.963	.927	.983	.965
		S3	.971	.944	.974	.949	.977	.955	.973	.947	.973	.947	.982	.965
		S4	.967	.935	.969	.938	.970	.941	.966	.934	.966	.934	.976	.952
	RF -3	S1	.977	.954	.977	.955	.985	.970	.976	.952	.977	.954	.989	.978
		S2	.970	.941	.973	.947	.983	.966	.969	.938	.968	.937	.984	.968
		S3	.976	.953	.977	.955	.981	.962	.977	.954	.977	.954	.983	.965
		S4	.970	.941	.970	.942	.978	.956	.969	.939	.969	.939	.982	.964
Δ			.035	.067	.033	.063	.102	.191	.034	.064	.034	.065	.043	.083
Mean			.963	.928	.965	.932	.961	.924	.964	.929	.963	.928	.975	.950
D-3	RF -1	S1	.953	.909	.952	.907	.895	.801	.951	.904	.939	.882	.955	.912
		S2	.928	.861	.920	.847	.863	.744	.920	.847	.659	.435	.936	.876
		S3	.936	.877	.934	.872	.947	.897	.935	.874	.930	.864	.958	.918
		S4	.942	.887	.941	.885	.945	.894	.940	.883	.938	.880	.952	.906
	RF -2	S1	.961	.924	.960	.922	.961	.924	.965	.931	.751	.565	.975	.950
		S2	.958	.917	.959	.920	.968	.937	.959	.919	.961	.924	.972	.945
		S3	.948	.898	.944	.890	.955	.912	.946	.896	.952	.906	.968	.937
		S4	.961	.923	.958	.918	.969	.939	.960	.922	.961	.924	.975	.951
	RF -3	S1	.972	.945	.973	.947	.977	.954	.975	.950	.974	.948	.973	.946
		S2	.968	.937	.968	.937	.973	.946	.966	.933	.961	.924	.967	.934
		S3	.950	.902	.949	.901	.972	.944	.950	.903	.948	.898	.972	.945
		S4	.955	.912	.953	.908	.973	.947	.953	.907	.951	.905	.981	.963
Δ			.044	.084	.053	.100	.114	.210	.055	.103	.315	.513	.045	.087
Mean			.953	.908	.951	.905	.950	.903	.952	.906	.910	.838	.965	.932

D: Distribution type, RF: Response format, S: Sample Size

Table 3. Correlation and determination coefficients for situation-3

D	RF	S	Situation-3											
			CTT		PCA		MR		ULS		DWLS		GRM	
			R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
D-1	RF -1	S1	.837	.701	.813	.660	.808	.653	.799	.639	.805	.648	.877	.769
		S2	.889	.790	.885	.783	.915	.837	.888	.789	.882	.778	.926	.857
		S3	.908	.824	.904	.817	.932	.869	.899	.809	.902	.814	.936	.877
		S4	.886	.785	.879	.773	.913	.834	.881	.777	.884	.781	.915	.838
	RF -2	S1	.907	.823	.902	.814	.951	.905	.901	.811	.899	.808	.957	.916
		S2	.914	.835	.911	.829	.941	.886	.909	.826	.909	.826	.954	.910
		S3	.907	.823	.902	.814	.948	.898	.900	.811	.899	.808	.948	.899
		S4	.918	.842	.910	.829	.953	.908	.909	.826	.903	.816	.953	.909
	RF -3	S1	.915	.837	.911	.829	.963	.927	.910	.829	.752	.566	.956	.914
		S2	.933	.870	.932	.868	.961	.924	.932	.869	.932	.868	.962	.925
		S3	.901	.812	.899	.809	.959	.919	.898	.806	.898	.807	.954	.910
		S4	.884	.781	.882	.778	.961	.924	.880	.774	.879	.773	.953	.907
Δ			.096	.169	.119	.208	.155	.274	.133	.230	.180	.302	.085	.156
Mean			.900	.810	.894	.800	.934	.874	.892	.797	.879	.774	.941	.886
D-2	RF -1	S1	.822	.676	.814	.662	.777	.604	.811	.658	.758	.574	.847	.717
		S2	.819	.671	.800	.640	.803	.644	.797	.636	.798	.637	.868	.753
		S3	.789	.622	.777	.604	.833	.693	.773	.598	.774	.599	.863	.745
		S4	.816	.666	.801	.642	.866	.751	.798	.637	.798	.637	.875	.766
	RF -2	S1	.765	.585	.737	.543	.836	.699	.735	.541	.657	.431	.777	.603
		S2	.811	.658	.794	.631	.875	.765	.790	.625	.790	.624	.750	.562
		S3	.804	.646	.794	.631	.882	.779	.790	.624	.790	.625	.860	.740
		S4	.828	.685	.802	.643	.894	.799	.800	.641	.799	.639	.900	.810
	RF -3	S1	.781	.610	.757	.573	.905	.818	.737	.544	.624	.389	.706	.498
		S2	.818	.669	.808	.652	.911	.829	.791	.626	.795	.632	.896	.803
		S3	.783	.614	.773	.598	.906	.820	.771	.595	.768	.590	.893	.797
		S4	.793	.629	.780	.609	.909	.826	.774	.599	.773	.597	.880	.775
Δ			.063	.100	.077	.119	.134	.225	.076	.117	.175	.250	.194	.312
Mean			.802	.644	.786	.619	.866	.752	.781	.610	.760	.581	.843	.714
D-3	RF -1	S1	.664	.441	.595	.354	.559	.313	.536	.288	.040	.002	.517	.267
		S2	.746	.557	.707	.500	.670	.448	.682	.466	.612	.374	.754	.568
		S3	.727	.529	.670	.449	.646	.418	.667	.444	.669	.448	.734	.538
		S4	.734	.538	.676	.457	.653	.427	.678	.460	.677	.459	.791	.625
	RF -2	S1	.684	.467	.572	.328	.550	.303	.538	.290	.587	.344	.525	.275
		S2	.698	.488	.656	.431	.631	.399	.641	.411	.545	.297	.685	.469
		S3	.705	.497	.623	.388	.737	.544	.613	.376	.637	.406	.713	.508
		S4	.668	.446	.637	.406	.741	.549	.628	.395	.622	.386	.741	.550
	RF -3	S1	.755	.571	.734	.539	.715	.512	.709	.503	.596	.355	.634	.402
		S2	.677	.458	.636	.405	.708	.501	.612	.374	.625	.391	.672	.452
		S3	.721	.520	.695	.483	.684	.468	.691	.478	.695	.483	.770	.593
		S4	.668	.446	.633	.400	.764	.584	.614	.377	.613	.376	.733	.538
Δ			.091	.130	.162	.211	.214	.281	.173	.215	.655	.481	.274	.358
Mean			.704	.497	.653	.428	.672	.456	.634	.405	.577	.360	.689	.482

D: Distribution type, RF: Response format, S: Sample Size

The coefficients of relationship obtained from the negatively skewed distribution in Situation-3 were high (r>.80). The correlation coefficients for the parameter estimates that the MR and DWLS models yielded increased particularly as the sample sizes increased. The average scores of the correlation coefficients that GRM yielded were the highest. The correlation coefficients obtained from the normal distribution in Situation-3 were moderate or high. The correlation coefficients that the DWLS and GRM models yielded were moderate in small sample sizes, but increased as the sample size increased. The correlation coefficients

averages obtained from MR were the highest. It was found that there was zero correlation between the true score and sample size-1 and response format-1 conditions of the DWLS model in the positively skewed distribution in Situation-3. A relationship of moderate degree was observed in the other simulation conditions. It was found that the correlation coefficients that the DWLS and GRM models yielded were affected more by the simulation conditions; the correlation coefficients that CTT yielded had the highest average scores.

Whether or not the determination coefficients were affected by different simulation conditions were analyzed by Factorial ANOVA. Separate analyses were run for each Situation. It was found that the distribution types for Situation-1 ($F(2, 215)=41.28$, $p<.001$) and the interaction of the distribution types and statistical model effect were significant ($F(10, 215)=4.60$, $p<.01$). The effects of the response formats ($F(2, 215)=1.24$, $p=.633$), the sample size ($F(3, 215)=1.30$, $p=.534$) and the model ($F(5, 215)=.68$, $p=.655$) on the determination coefficient was not found to be statistically significant. According to the Bonferroni test, to determine the significance of the distribution type effects, the determination coefficients obtained from a negatively skewed distribution were found to be significantly lower than those obtained from the normal and positively skewed distributions; the determination coefficients obtained from a normal distribution were significantly lower than those obtained from a positively skewed distribution.

It was found that the effect of the response formats ($F(2, 215)=27.59$, $p<.01$) and the interaction of the response formats and model ($F(10, 215)=2.01$, $p<.05$) in Situation-2 were statistically significant. No statistical significance was found regarding the effects of the distribution types ($F(2, 215)=11.75$, $p=.080$), the sample size ($F(3, 215)=1.65$, $p=.416$) and the model ($F(5, 215) = 1.77$, $p=.220$) on the determination coefficient. According to the findings of the Bonferroni test, the determination coefficients obtained from the datasets that included items scored across seven categories were higher when compared to those items scored across three or five categories.

In Situation-3, the effects of the distribution types ($F(2, 215)=156.31$, $p<.001$) and the model ($F(5, 215)=4.00$, $p<.01$), the interaction of the distribution types and the model ($F(10, 215)=4.94$, $p<.01$), the interaction of the response formats and the model ($F(10, 215)=4.55$, $p<.05$) and the interaction of the sample size and the model ($F(15, 215)=4.84$, $p<.01$) were found to be statistically significant. It was found that the effects of the response format ($F(2, 215)=.85$, $p=.502$) and the sample size ($F(3, 215)=11.36$, $p=.152$) on the determination coefficient were not statistically significant. When the Bonferroni test was administered based on the distribution types, the determination coefficient findings obtained from the negatively skewed distribution were found to be significantly higher than those obtained from the normal and the positively skewed distributions. Similarly, the determination coefficients obtained from the normal distribution were significantly higher than those obtained from the positively skewed distribution. Based on the model, it was found that CTT yielded higher determination coefficients than did the ULS and DWLS models; PCA yielded higher determination coefficients than did the DWLS model, and the MR and GRM models yielded higher determination coefficients than did the CTT, PCA, ULS and DWLS models.

4. DISCUSSION AND CONCLUSION

In the present research study, where the basic simulative conditions were an item difficulty level of 20% below average, 20% above average, and normal, various distribution types, the effects of such simulative conditions as response formats and sample sizes on estimating the latent ability distribution were also investigated. To this end, ability parameters of true latent traits were identified and latent trait estimates were made with six different models within related simulative conditions.

In Situation-1, when the item difficulty was low, the distribution was negatively skewed, the response format was three and the sample size was small, all the models yielded values that were not related to the true ability parameters. It is recommended that none of the models should be utilized under these simulative conditions. As the sample size and response categories increased, moderate relationships started to be observed. The MR model, low item difficulty level, and a negatively skewed distribution do not yield accurate parameter estimations; however, in normal distributions, the MR model displays a better performance than do all the other models. All the models, primarily the MR model, are affected more by the negatively skewed distribution and, thus, do not make accurate estimations. However, when compared to normal distributions, positively skewed distributions can be said to yield better findings. Under these simulative conditions, CTT, MR and GRM display the best performances.

In Situation-2, the estimations yielded by the MR model was found to be affected by negatively skewed distributions, especially when the sample size is small. In Situation-2, determination coefficients increase as the response format and sample size increase. Under these simulative conditions, the GRM model displays the best performance.

The coefficients of relationship obtained in Situation-3 were moderate or high. The relationship coefficients that the DWLS and GRM models yielded were found to be moderate when the sample size was small, but higher when the sample size increased. Under these simulative conditions, CTT, MR and GRM displayed the best performances.

The findings of ANOVA, which was administered to determine whether or not simulative conditions affected determination coefficients, showed that particularly distribution types had a significant effect on determination coefficients in negatively skewed and positively skewed distributions. In the present research, where the distribution of item difficulty levels and distribution types were both studied, a significant effect of distribution types was an expected findings. It was found that the response format in Situation-2 and the model in Situation-3 were simulative conditions that had a significant effect. This significant effect in Situation-3 was in favor of particularly GRM and MR. While in Situation-1 and Situation-2 the model did not have a significant effect, the average determination coefficient values of the MR and GRM models were higher than those yielded by the other models. This situation shows that the general performance levels of MR and GRM, which produced latent ability estimations, are high.

In Situation-2, it was found that the significant effect of the response format on the determination coefficient was in favor of a seven-category response format. This finding is consistent with those reported in studies by Allahyari, Jafari and Bagheri (2016) and by Lozano et al. (2008). Allahyari et al. (2016) reported in their study that particularly in situations where the potential distribution was not normal, increasing a three or five-category response format to a higher category level would increase the power of the statistical model of Differential Item Functioning (DIF) by 8%.

The finding that the ability parameters that GRM yielded were higher than almost all other models under different conditions showed consistency with the findings reported in studies by Dumenci and Achenbach (2008) and by Hauck Filho et al., (2014).

When the general performance of the models are evaluated, it can be said that MR and GRM display a better performance than the other models. Particularly in situations when the distribution is either negatively or positively skewed and when the sample size is small, these models display a rather good performance.

The present study can be further developed by means of further studies on different simulation conditions. Iterative and bayesian parameter estimations, such as particularly

Markov Chain Monte Carlo, can be used. In addition, this study, the structure of which was based on a single dimension, can be developed by using multidimensional structures. Moreover, different polytomous parameter estimation models of IRT (such as the rating scale model –RSM) or nonparametric item response theory models can be used.

ORCID

Hakan Kogar https://orcid.org/0000-0001-5749-9824

5. REFERENCES

Allahyari, E., Jafari, P., & Bagheri, Z. (2016). A simulation study to assess the effect of the number of response categories on the power of ordinal logistic regression for differential ıtem functioning analysis in rating scales. Computational and mathematical methods in medicine, vol. 2016, Article ID 5080826. doi.org/10.1155/2016/5080826

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach (Vol. 904). John Wiley & Sons. doi.org/10.1002/9781119970583

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. Intelligence, 30(6), 505-514. doi.org/10.1016/S0160-2896(02)00082-X

Brzezińska, J. (2016). Latent variable modelling and item response theory analyses in marketing research. Folia Oeconomica Stetinensia, 16(2), 163-174. doi.org/10.1515/foli-2016-0032

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cyr, A., & Davies, A. (2005). Item response theory and latent variable modeling for surveys with complex sampling design: The case of the national longitudinal survey of children and youth in Canada. In conference of the Federal Committee on Statistical Methodology, Office of Management and Budget, Arlington, VA.

Dawber, T., Rogers, W. T., & Carbonaro, M. (2009). Robustness of Lord's formulas for item difficulty and discrimination conversions between classical and item response theory models. Alberta Journal of Educational Research, 55(4), 512.

DeCoster, J. (1998). Overview of factor analysis. Retrieved June 12, 2017 from http://www.stat-help.com/factor.pdf

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. Psychological assessment, 20(1), 55-62. doi.org/10.1037/1040-3590.20.1.55

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. Applied Psychological Measurement, 31(5), 457-459. doi.org/10.1177/0146621607299271

Hauck Filho, N., Machado, W. D. L., & Damásio, B. F. (2014). Effects of statistical models and items difficulties on making trait-level inferences: A simulation study. Psicologia: Reflexão e Crítica, 27(4), 670-678. doi.org/10.1590/1678-7153.201427407

Jafari, P., Bagheri, Z., Ayatollahi, S. M. T., & Soltani, Z. (2012). Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQL TM 4.0 Generic Core Scales in school children. Health and Quality of Life Outcomes, 10(1), 27. doi.org/10.1186/1477-7525-10-27

Kline, R. B. (2005). Principles and practice of structural equation modeling (Second Edition). New York: The Guilford Publications.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behavior Research Methods, 48(3), 936-949. doi.org/10.3758/s13428-015-0619-7

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. Behavior research methods, 38(1), 88-91. doi.org/10.3758/BF03192753

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. Methodology, 4(2), 73-79. doi.org/10.1027/1614-2241.4.2.73

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. Behavior Research Methods, 41(2), 295-308. doi.org/10.3758/BRM.41.2.295

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. Psychological Methods, 1, 293 – 299. doi.org/10.1037/1082-989X.1.3.293

Raykov, T. ve Marcoulides, G. A. (2000). *A first course in structural equation modeling.* London: Lawrence Erlbaum Associates, Inc.

Revelle, W. (2017). Package 'psych'. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rizopoulos, D. (2017). Package 'ltm'. Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monographs, 34(Suppl. 17).

Saporta, G., & Niang, N. (2009). Principal component analysis: Application to statistical process control. Data analysis, 1-23. doi.org/10.1002/9780470611777.ch1