Investigating the Effect of Item Order on the Psychometric Properties of a Self-Efficacy Perception Scale

Assist. Prof. Dr. Elif Kübra Demir
Ege University- Turkiye
ORCID: 0000-0002-3219-1644
elif.kubra.demir@ege.edu.tr

Abstract

In this study, the item order effect was investigated using the self-efficacy perception scale for computational thinking. 946 participants consisting of 8th-grade students participated in the research. Participants were first administered the original form of the scale, in which the dimensions were sequential and each item was presented in the relevant subdimension. One month later, a second form in which the order of the items was completely random was administered to the same group. Analyzes revealed no significant difference in the group's total mean score between the two forms. Additionally, the study showed the extent to each individual participant's total and factor scores varied between the two forms, and the difference was negligible. Another remarkable finding is that sequentially ordering items representing the same dimension contributes positively to the internal consistency reliability of the scale. Confirmatory factor analyses were performed for both forms, revealing good model fit with similar index values; This shows that randomizing the item order does not disrupt the structure of the scale. In addition, factor loading differences between the models were also examined. Finally, measurement invariance was examined and it was found that there was a solid level of measurement invariance between the two forms. In future studies, instead of creating random forms, forms in which the order of the items is consciously manipulated can be used and thus contextual effects can be examined. It is also suggested that the item order effect can be examined in the context of many demographic characteristics and different item types.

Keywords: *Item order, Item position, Self-report scale, Measurement invariance, Self-efficacy*

E-International Journal of Educational Research

Vol: 14, No: 5, pp. 478-493

Research Article

Received: 2023-09-18
Accepted: 2023-10-17

Suggested Citation

Demir, E. K. (2023). Investigating the Effect of Item Order on the Psychometric Properties of a Self-Efficacy Perception Scale, *E-International Journal of Educational Research,* 14 (5), 478-493. DOI: https://doi.org/10.19160/e-ijer.1362442

INTRODUCTION

There are various error sources or factors that can influence individuals' test scores. These factors generally stem from the measurement tool itself, the conditions under which the measurement is conducted, or the characteristics of the test takers. When the error variance of these types of factors is introduced into measurement results, the desired true performance results may not be achieved. Factors such as individuals' mental states like anxiety, sadness, happiness at the time of taking the test, societal influences like social desirability, and physiological effects like fatigue and sleep deprivation can cause test scores to deviate from true performance scores. In addition to many uncontrollable factors, the problems that test developers can solve are errors caused by the measurement tool itself. Essentially, the stimuli given to individuals in a measurement tool are test items, and these items, alone or in interaction with other items, affect the test scores of individuals. There are many variables that can cause errors in the items, such as the suitability of the items for the targeted age group, the complexity of the expressions, and the appropriateness of the response category (Cronbach, 1988, 2013, Haladayna & Rodriguez, 2013; Osterlind, 1998). One of the most important of these is the problem of whether the order of items in a test has an effect on response behavior and therefore on the scores obtained.

The item order or context effects refer to the idea that earlier items in a test may influence later responses (Schuman & Presser, 1981). The pioneering studies that preceded the Item-Order Effect initially focused on the impact of stimuli presented to individuals, such as information, words, etc., rather than the order of items, on their response behavior. The first of these studies can be attributed to Lund (1925), whose research examined how the presentation of two conflicting pieces of information to individuals influenced their attitudes. Lund's study indicated that controversial topics had a greater impact on individuals' attitudes compared to uncontroversial ones (Crano, 1977). Thus, the issue of primacy-recency effect comes into play when collecting information about individuals (Chen, 2010). Bossart and Di Vesta (1966) used adjective strings to have university students rate their impressions of less familiar individuals and obtained the result that student impressions of the individuals tended to be more positive when positive adjectives were presented first. Bradburn (1983) emphasizes the importance of understanding how the order can influence the results of a scale, despite there being little definitive evidence regarding a specific item order effect.

Brenner (1964) conducted an achievement test by changing the item order to examine the reliability, difficulty, and discrimination of the test. The results showed that there was no statistically significant difference in terms of test difficulty, discrimination, and test reliability when the items were arranged from easy to difficult, difficult to easy, or randomly. Similarly, Munz and Smouse (1968) observed students' achievement test scores with three different item difficulty patterns, namely easy to difficult, difficult to easy, and random, and found that item difficulty order had no effect on the total test score. Klosner and Gellman (1973) compared students' test performance with three test types. In the tests, the order of items is arranged according to topics, from easy to difficult within topics, and from easy to difficult between topics. There was no significant difference between test scores. Hambleton and Traub (1974) examined the mathematics test performance of high school students by ordering the items from easy to difficult and from difficult to easy, and found that students obtained higher scores on items ordered from easy to difficult. Perlini et al. (1988) investigated the contextual effects in exams. They examined how factors like exam duration, item order, chapter order, and difficulty level influence test performance. The results showed that contextual factors, particularly exam duration, had an impact on overall performance. However, factors like item order, chapter, and difficulty arrangements had minimal to no effect on overall performance. Substantive individual differences in learning effects were found within tests measuring fluid intelligence, such as the Advanced Progressive Matrices (APM) (Schweizer et al. 2009). In another study, it was emphasized that using test forms with altered item order had no significant effect on response behavior, or that any effect could be considered negligible (Albano, 2013; Asseburg & Frey, 2013). Demirkol and Kelecioğlu (2022), investigated the effect of item position in the fields of reading and mathematics on the PISA 2015 Turkey sample with explanatory item response theory. The results indicate that the item position effect led to a reduced likelihood of providing correct answers, particularly in the realm of reading as opposed to mathematics. Furthermore, within the domain

of mathematics, open-response items were more influenced by item position compared to multiple-choice items.

Strack et al. (1988) emphasized that the order of items in attitude scales can hold significant importance in terms of the obtained results. However, Knowles (1988) investigated the item order effect in personality tests. Multiple forms were created, allowing each item to be presented at every possible position throughout the measurement. The findings indicated that the average score was not affected by the item order. In a study conducted by Vega and O'Leary (2006) with 641 students to examine partner aggression, they administered two forms of the test in which they manipulated the item order. They concluded that the test results were not affected by the item arrangement. Chen (2010) conducted a study to investigate the effects of item order on attitude measurements related to item difficulty, item discriminability, test scores, test length, response time, and test reliability. It is found that the item order had an impact on attitude measurements, indicating that participants' responses to subsequent items could vary depending on the items presented earlier. Weinberg et al. (2018) conducted a study to evaluate the item order effect from a psychometric validation perspective. They created two forms, one with fixed domain items and general items randomly, the other with random domain items and fixed general items, and applied these forms to two different groups. The mean values obtained with fixed field forms are significantly higher than those obtained with random field forms, and both forms represent the structure similarly. As a result of the research, it is recommended to examine the item order effect of all new scales in order to reduce measurement error and increase accuracy in psychological evaluation. Şahin (2021) used the cyberloafing scale in his research to examine the item order effect by presenting a fixed-order form in which all items belonging to the same dimension were presented together and a random-order form in which the items were randomly ordered. Results from two separate groups drawn from the sample revealed a statistically significant difference between the mean scores of the two forms. Multi-group CFA (mg-CFA) was performed to determine whether measurement invariance was achieved with different item order presentations of the scale. The findings indicate that measurement invariance could not be achieved even at the first stage of the analysis.

Examining studies on the item order effect reveals two important problems. First of all, the majority of studies focused on the item order effect in maximum performance tests and were generally associated with item difficulty. Item order effects on measures targeting typical performance, such as attitudes, interests, and perceived competence, have not been adequately investigated. There are conflicting results in studies regarding both types of testing. It is still difficult to find a clear answer regarding item order effects in the literature, and this is an important problem in terms of the development and application of both psychological and achievement tests. In typical performance-oriented scales, this information is even more limited, so the item order effect needs to be examined in such scales. Therefore, the purpose of this study was to investigate item order effects using a computational thinking self-efficacy perception scale called the Self-Efficacy Perception Scale for Computational Thinking Skills (CTSSP), which measures typical performance. Another problem is that studies collecting data from the same group are very limited and most of the studies were conducted with university students. Therefore, in order to contribute to the solution of this problem, the participants of this study were 8th grade secondary school students and the item order effect was examined in the same group. Answers were sought to the questions of whether different item orders caused a significant change in the psychometric properties of the scale and the average score of the group, the amount of score difference between each individual's scores from both forms, whether the scale was consistent or not, and whether the same structure and measurement invariance were achieved.

METHOD

1. Participants

This study was conducted with middle school 8th-grade students and comprised two data collection phases. Initially, the original version of the scale was administered to the students. In the original-order form, the items on the scale were arranged under their respective dimensions, which consisted of five subdimensions. In the random-order form of the scale, the items were randomly

ordered without considering their dimensions. The students were first administered the original-order form of the scale, followed by the random-order version one month later. In the initial phase, 1088 students participated. 64 students were unable to participate in both applications, and 78 students were excluded from the study who did not provide responses more than half of the items. Consequently, the study was conducted with 946 students, of whom 51% were female, and 49% were male.

2. Data Collection Tools and Procedures

In this study, self-efficacy perception scale for computational thinking skill (CTSSP), developed by Gülbahar et al. (2019), was utilized. The CTSSP was specifically designed for secondary school students. To ensure the content and face validity of the scale, a scale development study was conducted based on expert opinions, involving 3 domain experts, 4 subject teachers, and a Turkish language specialist. The response categories were designed in line with expert recommendations and took into consideration age group characteristics, resulting in a three-point graded scale (1-Yes, 2-Partly, 3-No). Exploratory and confirmatory factor analyses were conducted using data obtained from 916 participants, and internal consistency reliability coefficients were calculated. The analyses led to the establishment of a five-dimensional scale structure consisting of 36 items. These dimensions are named as Algorithm Design Competency (9 items), Problem Solving Competency (10 items), Data Processing Competency (7 items), Basic Programming Competency (5 items), and Self Confidence Competency (5 items).

Within the original form of the scale, items corresponding to each of these five dimensions are presented collectively under their respective dimension. To measure the item order effect, items were randomly rearranged, creating another form referred to as the random form. Initially, the original form was administered to all students, followed by the application of the random form in the second session. In studies of this nature, the elapsed time between applications holds paramount significance. The principle applied in determining the duration was to ensure that it is brief enough for the measured attribute to remain relatively stable, yet long enough for items not to be readily recalled (Fraenkel & Wallen, 2009). For the affective attribute of self-efficacy perception, a one-month interval between the two applications was established. Students who did not participate in the initial session were not included in the second session. Moreover, 64 students who participated in the first session but not in the second were excluded from the study. Additionally, 78 students who failed to respond to more than half of the items in the scales were removed from the research dataset.

3. Data Analysis

In order to test whether there was a significant difference between the mean scores of students for the original-order form and random-order form, a paired-samples t-test was employed. Also Pearson's product-moment correlation coefficient between the scores from the original form and the random form examined. Confirmatory factor analysis (CFA) was conducted for each form, examining the structure of the scales, factor loadings, and model fit indices. Additionally, in this study, McDonald's ω coefficients was calculated to evaluate composite reliability, Cronbach's α and Stratified α coefficient were calculated to measure internal consistency. To assess potential measurement invariance across scale forms, a multiple-group CFA analysis was performed. Measurement invariance is a crucial assessment that examines whether a construct maintains its psychometric properties consistently across different groups or measurement occasions. It verifies that the construct holds the same meaning for various groups or over multiple measurements (Bornstein, 1995; Glanville & Wildhagen, 2007; Little, 2013; Widaman et al., 2010; Putnick & Bornstein, 2016). The test of measurement invariance consists of a four-stage model (Horn & McArdle, 1992; Meredith, 1993; Widaman & Reise, 1997). The four stages are expressed as configural, metric, scalar, and strict invariance, starting with the least constrained model and moving toward the most constrained model. Each model is analyzed in comparison with the previous one. Since strict invariance is the exact equivalence of the measurement model, it has been considered very difficult to achieve this in practice. Scalar invariance remains the widely accepted standard of measurement invariance in psychology, allowing the use of observed scores (Luong & Flake, 2022). Also since strict invariance does not have any direct consequences for the comparability of structural parameters across groups, most applied studies refrain from its evaluation (Leitgöb et al., 2022; Putnick & Bornstein, 2016; Şahin, 2021). However, in this study, the analysis was continued up to the

strict invariance step, which is the most difficult level of measurement invariance to achieve. Chi-square difference tests and analyses of the CFI change between models were utilized to assess measurement invariance. The finding that the chi-square change value was p>.05 and CFI ≤ 0.01 indicates that measurement invariance was established at that model level (Cheung & Rensvold, 2002; Hu & Bentler, 1998). Analyzes were conducted with Mplus 8.0 (Muthén & Muthén, 2017).

4. Ethics of the study

Before starting the research, the authors who developed the scale were informed about the study, and their permission was obtained. During the application phase of the scale, consent forms were taken from the students and only volunteer students participated in the study. No personal or demographic information, including name information, was collected. Participants filled out the forms by choosing a nickname, and the forms in both separate applications were matched with this nickname.

FINDINGS

In order to examine the relationships between the scores obtained from the original-order form and the random-order form, in terms of both the total score and the subdimensions, the Pearson correlation coefficient was calculated and presented in Table 1. There is a very high positive correlation (r=0.98) between the total scores of forms with different item orders. Similarly, in terms of subdimensions, the correlation coefficient between both forms varies between 0.79 and 0.72.

Table 1. Pearson correlation between scale scores of the forms

Variable	Correlation coefficient
Total Scale	0.98
Algorithm Design Competency	0.78
Problem Solving Competency	0.72
Data Processing Competency	0.79
Basic Programming Competency	0.74
Self Confidence Competency	0.74

Since a high positive correlation between total scores does not mean that there is no significant difference between the scores, the descriptive statistics for the scores obtained from both the original and random forms of the CTSSP scale were analyzed, and paired t-test results are delineated in Table 2. The table indicates that there was no statistically significant difference, t(945)= 1.67, p=0.10, between the total scale mean scores of participants who took the original order form (M = 80.95, SD = 14.27) and those who took the random order form (M = 80.83, SD = 14.35). Similarly, there is no significant difference between the original-ordered form and the random-ordered form for all subdimensions (p>.05).

Table 2. *T-test results for original and random ordered forms*

	Original-Order		Random-Order				
Variable	***M***	***SD***	***M***	***SD***	***t***	***df***	***p***
Total Scale	80.95	14.27	80.83	14.35	1.67	945	0.10
Algorithm Design	17.24	5.70	17.21	5.57	0.14	945	0.88
Problem Solving	24.33	4.04	24.35	4.28	-0.11	945	0.91
Data Processing	16.58	3.72	16.56	3.85	0.21	945	0.83
Basic Programming	10.51	3.07	10.50	3.04	0.09	945	0.93
Self Confidence	12.28	2.48	12.24	2.38	0.55	945	0.58

The mean differences between the scores obtained by each student from the original-order form and the random-order form have been calculated. For the subdimensions, these difference scores are calculated as 2.25, 2.69, 2.35, 1.98, and 2.05, respectively. The average difference between the total scores obtained by each student from the two forms is found to be 0.89.

To examine the reliability in terms of internal consistency, Cronbach's α and McDonald's ω, Stratified α coefficients were calculated for both forms. Stratified α coefficient was used for the entire scale and was calculated as 0.94 for the original-order form and 0.93 for the random-order form. Cronbach's α and McDonald's ω coefficients calculated for the subdimensions are given in Table 3. When the table is examined, it is seen that the coefficients are quite close to each other. While Cronbach's α coefficients vary between 0.92 and 0.75 for the Original-order form, they vary between 0.90 and 0.79 for the Random-order form. It can be stated that the scores obtained from the Scale for both forms provide internal consistency and are reliable. However, when comparing the forms, it can be noted that the Stratified α coefficient is slightly higher for the original-order form. Similarly, for all subdimensions Cronbach's α coefficient is slightly higher in the original-order form.

Table 3. *Reliability coefficients for subdimensions*

		Original-Order		Random-Order	
Subdimension	**Number of items**	**Cronbach's α**	**McDonald's ω**	**Cronbach's α**	**McDonald's ω**
Algorithm Design	9	0.92	0.92	0.90	0.92
Problem Solving	10	0.86	0.85	0.85	0.85
Data Processing	7	0.88	0.88	0.86	0.88
Basic Programming	5	0.86	0.86	0.84	0.86
Self Confidence	5	0.77	0.78	0.75	0.75

The univariate normality assumption was confirmed, but the multivariate normality assumption was violated. Two separate CFAs for the original ordinal and random order forms were used with the MLR estimator. Model fit statistics for these two forms are given in Table 4. In terms of model fit statistics, SRMR and RMSEA values being less than 0.08 and χ2/sd value being less than 5 indicate that the model data fit is good (Jöreskog & Sörbom 1993). Another of the most commonly used cut-off values for RMSEA is stated by Hu and Bentler (1999), as values lower than 0.06 indicating good fit. For CFI, values of 0.90 and above are defined as acceptable (McDonald & Ho, 2002). Considering these cut-off points, it can be stated that model fit is achieved for both forms. In terms of CFI and TLI indices, model fit values of both forms are the same. However, considering the χ2/df ratio, SRMR, and RMSEA values, the model fit of the original order form is slightly better than that of the random-order form.

Table 4. *CFA model fit indices*

Form	**χ2**	**df**	**CFI**	**TLI**	**SRMR**	**RMSEA**	**%90 C.I. RMSEA**
Original-order	1949.55	584	0.91	0.89	0.047	0.050	0.047 ; 0.052
Random-order	1994.82	584	0.91	0.89	0.053	0.051	0.048 ; 0.053

Factor loadings, residual errors, mean and standard deviation values of the items obtained as a result of CFA of both models are given in Table 5. When the factor loadings of the items are examined, it is seen that the factor loadings are surprisingly the same in the first four dimensions. Factor loadings in these dimensions vary between 0.81 and 0.51. In the Self Confidence Competency subdimension, the values of the original-order form are higher in all factor loadings. The lowest value among all factor loadings is in random-order form with a value of 0.47. The factor loading of the same item in the original-order form was calculated as 0.53.

Table 5. *CFA values*

	Original-order				Random-order			
Subdimension	***M***	***SD***	***λ***	**S.E**	***M***	***SD***	***λ***	**S.E.**
Algorithm Design								
Item 1	2.12	0.81	0.77	0.402	2.14	0.80	0.77	0.40
Item 2	2.17	0.83	0.74	0.447	2.14	0.82	0.74	0.44
Item 3	2.03	0.83	0.80	0.355	2.04	0.84	0.80	0.35
Item 4	1.76	0.76	0.79	0.378	1.77	0.75	0.79	0.38
Item 5	1.73	0.78	0.72	0.477	1.73	0.76	0.72	0.47
Item 6	2.01	0.82	0.81	0.343	1.98	0.80	0.81	0.34
Item 7	1.90	0.81	0.78	0.399	1.89	0.77	0.78	0.39
Item 8	1.88	0.78	0.80	0.364	1.88	0.77	0.80	0.36
Item 9	1.65	0.77	0.63	0.601	1.64	0.74	0.63	0.60
Problem Solving								
Item 10	2.55	0.62	0.61	0.632	2.54	0.63	0.61	0.63
Item 11	2.58	0.62	0.64	0.588	2.58	0.62	0.64	0.58
Item 12	2.61	0.60	0.64	0.592	2.57	0.61	0.64	0.58
Item 13	2.56	0.63	0.62	0.621	2.55	0.62	0.62	0.60
Item 14	2.44	0.64	0.66	0.570	2.42	0.65	0.66	0.57
Item 15	2.36	0.67	0.57	0.678	2.38	0.68	0.57	0.58
Item 16	2.45	0.67	0.55	0.696	2.43	0.67	0.55	0.65
Item 17	2.15	0.73	0.51	0.736	2.23	0.71	0.51	0.62
Item 18	2.33	0.70	0.59	0.653	2.35	0.68	0.59	0.63
Item 19	2.32	0.71	0.61	0.629	2.29	0.71	0.61	0.58
Data Processing								
Item 20	2.58	0.66	0.72	0.479	2.56	0.66	0.72	0.46
Item 21	2.46	0.69	0.78	0.397	2.47	0.70	0.78	0.39
Item 22	2.47	0.70	0.75	0.445	2.45	0.71	0.75	0.34
Item 23	2.22	0.72	0.69	0.530	2.22	0.72	0.69	0.43
Item 24	2.34	0.73	0.74	0.456	2.33	0.73	0.74	0.45
Item 25	2.34	0.73	0.74	0.456	2.34	0.73	0.74	0.45
Item 26	2.18	0.77	0.61	0.627	2.19	0.79	0.61	0.42
Basic Programming								
Item 27	2.29	0.74	0.65	0.573	2.29	0.74	0.65	0.47
Item 28	1.99	0.79	0.79	0.382	2.03	0.76	0.79	0.38
Item 29	1.98	0.80	0.81	0.341	2.00	0.78	0.82	0.34
Item 30	2.09	0.79	0.76	0.425	2.05	0.76	0.76	0.42
Item 31	2.15	0.80	0.69	0.526	2.13	0.77	0.69	0.50
Self Confidence								
Item 32	2.51	0.68	0.66	0.538	2.48	0.65	0.68	0.50
Item 33	2.47	0.67	0.69	0.547	2.45	0.66	0.67	0.42
Item 34	2.50	0.67	0.66	0.575	2.51	0.66	0.65	0.44
Item 35	2.44	0.68	0.65	0.596	2.44	0.66	0.63	0.52
Item 36	2.37	0.73	0.53	0.780	2.36	0.70	0.47	0.54

Chi-square, CFI, and RMSEA difference tests were used to examine measurement invariance. The fact that the chi-square change value is p>.05 and CFI, RMSEA ≤ 0.01, indicates that measurement invariance is achieved at the relevant level (Cheung & Rensvold, 2002; Hu & Bentler, 1998; Putnick & Bornstein, 2016). As seen in Table 6, all significance values of the chi-square difference test were calculated as p>.05. At the same time, for all measurement invariance levels ΔCFI and ΔRMSEA values

are less than 0.01. In this case, it can be stated that there is a strict level of measurement invariance between the original-order and random-order forms.

Table 6. *Measurement invariance fit indices*

Level	χ^2	*df*	$\Delta\chi^2$	*p*	CFI	ΔCFI	RMSEA	ΔRMSEA
Configural	3944.77	1168			0.903		0.054	
Metric	3991.55	1199	24.39	0.79	0.903	0.000	0.053	0.001
Scalar	4032.39	1230	29.22	0.56	0.902	0.001	0.052	0.001
Strict	4077.62	1266	32.34	0.64	0.901	0.001	0.052	0.000

CONCLUSION, DISCUSSION AND SUGGESTIONS

Studies related to the item order effect, which is believed to potentially influence measurement results, present conflicting findings in the literature. Upon reviewing the literature, it is observed that research conducted on characteristics such as attitude, interest, and ability within the same group is limited. The item order effect is investigated in this study by administering the CTSP scale, which measures competence perception, to the same group first in its original-order and then in a completely random-order, with a one-month interval.

In the original-order form used, the items were grouped under their respective subdimensions. The difference between the scores obtained between the two forms was examined, and no significant difference was found between the total scores. This result aligns with the research findings of Munz and Smouse (1968), Krampen et al. (1992), O'Shaughnessy (1983), Laffitte (1984), Solomon and Kopelman (1984), Vega and O'Leary (2006), Etzel et al. (2021), Hartig et al. (2007), Albano (2013), Asseburg and Frey (2013), while differing from the findings of Bossart and Di Vesta (1966), Flaugher et al. (1968), Hambleton and Traub (1974), Strack et al. (1988), Dillman (2000), and Şahin (2021).

The extent to which each participant's scores differed between the two administrations was examined, and the mean difference for the total score was calculated as 0.89. For the subdimensions, these differences range between 1.98 and 2.69. Due to the absence of a difference between the group means and the small mean difference between individual scores, it can be concluded that the item order has no significant effect on the scores. However, it should be noted that this study was conducted with a self-efficacy perception scale, and in measurements where small score differences are significant, this result may be interpreted differently for critical decisions. To determine the reliability of the scales, Cronbach's α and McDonald's ω, as well as Stratified α coefficients, were examined. It can be stated that the Stratified α value for the total scale is slightly higher in the original-order form, but very high coefficients were obtained for both forms. Cronbach's α coefficients for the subdimensions were examined, it was found that the values for the original-order form were higher for all subdimensions. As for McDonald's ω, the values for the first four subdimensions are the same, while for the last subdimension, the value for the original-order form is higher. Giving items representing the same dimension in the measurement tool in successive order contributed to the internal consistency reliability. Similarly, in his investigation using the cyberloafing scale, Şahin (2021) discovered greater internal consistency coefficients in the fixed-order form.

In order to examine the effect of item order on the structure of the scale and factor loadings, data obtained with each form were subjected to separate Confirmatory Factor Analyses (CFA). Both models demonstrated good model fit with similar model fit index values, indicating that randomizing the items did not disrupt the structure of the scale consisting of 5 subdimensions; the same structure was preserved. When examining the factor loadings, it was observed that the factor loadings were almost

identical for the first four factors. In the last subdimension named Self Confidence Competency, the factor loadings for the original-order form are higher. It can be stated that having items representing the same factor together has a positive effect on the factor loadings, but only for this subdimension. Multi-group CFA analysis was performed to test measurement invariance between forms. The analysis results indicate that measurement invariance between the two models is achieved at a strict level, which is a level that is difficult to attain in practice. Therefore, it can be stated that there is measurement invariance between giving the items in ordered or random form in the CTSP scale.

In the studies within the literature concerning the item order effect, it is observed that different and contradictory results have been obtained. In fact, among some studies, it can be seen that some of the variables such as mean score change, preservation of scale structure, internal consistency, and measurement invariance yield consistent results, while others produce contradictory results. This situation may be attributed to the use of different measurement tools in the studies. In studies where maximum performance is measured, the difficulty level of the items becomes more important, while in tests measuring typical performance, contextual effects between items become more significant. Item order effect can yield different results in these two types of tests. Different item-order effects can also be observed in scales measuring different characteristics such as attitude, interest, and competency perception. In this study, a measurement tool related to computational thinking self-efficacy perception was used. Self-efficacy perception may be a characteristic less affected by item-order effect. Therefore, the importance of the item-order effect may vary depending on the measured characteristic. In future research, studies focusing on the differentiation of the measured characteristic, especially in measuring typical performance, can be conducted. In attitude scales, contextual effects of items may be more observable. To investigate this effect, forms where the sequence is manipulated for a specific purpose rather than given randomly can be used. For example, negative items can be clustered together to examine their effects on item responses. In this study, the original-order and random-order forms were administered to the same group with a time interval, in future research, the forms can be cross-used in two separate groups to avoid the pretest effect. Additionally, the effect of item position can be examined in different sample sizes, age groups, demographic characteristics, and cultures. Research that can reveal the relationship between item type and item order effect will also make a significant contribution.

Madde Sıralamasının Bir Öz Yeterlik Algısı Ölçeğinin Psikometrik Özelliklerine Etkisinin İncelenmesi

Dr. Öğrt. Üyesi Elif Kübra Demir
Ege Üniversitesi - Türkiye
ORCID: 0000-0002-3219-1644
elif.kubra.demir@ege.edu.tr

Özet

Bu çalışmada bir öz yeterlik algısı ölçeği olan bilgi işlemsel düşünme becerisine yönelik öz yeterlik algısı ölçeği kullanılarak madde sıralama etkisi araştırılmıştır. Araştırmaya 8. sınıf öğrencilerinden oluşan 946 katılımcı katılmıştır. Katılımcılara öncelikle ölçeğin boyutlarının sıralı olduğu ve her maddenin ilgili alt boyutta sunulduğu orijinal form uygulanmıştır. Bir ay sonra aynı gruba madde sıralamasının tamamen rastgele olduğu ikinci bir form uygulanmıştır. Grubun bu iki forma ilişkin ölçeğin tamamı ve alt boyutlarından aldıkları ortalama puanları arasında anlamlı bir fark olmadığı bulunmuştur. Ayrıca, her bir katılımcının puanlarının iki form arasındaki değişimi incelenmiş ve bu farkın göz ardı edilebilir olduğu görülmüştür. Dikkat çeken bir diğer bulgu ise aynı boyutu temsil eden maddelerin birlikte sıralanmasının ölçeğin iç tutarlığına olumlu katkı sağlamasıdır. Her iki ölçek formu için de doğrulayıcı faktör analizleri yapılmış ve formlar birbirine yakın model uyum iyiliği göstermiştir. Bu durum madde sıralamasının rastgele yapılmasının ölçeğin yapısını bozmadığını göstermektedir. Son olarak ölçme değişmezliği incelenmiş ve iki form arasında katı düzeyde ölçme değişmezliği olduğu bulunmuştur. Gelecek çalışmalarda rastgele sıralanmış maddeler kullanmak yerine, madde sıralamasının belli bir bağlamsal amaca yönelik olarak manipüle edildiği formlar kullanılabilir. Ayrıca madde sıralama etkisinin farklı demografik özellikler ve farklı madde türü bağlamında incelenmesi de önerilmektedir.

Anahtar Kelimeler: *Madde sıralama etkisi, Madde konumu, Öz bildirim ölçeği, Ölçme değişmezliği, Öz yeterlik*

E-Uluslararası Eğitim Araştırmaları Dergisi

Cilt: 14, No: 5, ss. 478-493

Araştırma Makalesi

Gönderim: 2023-09-18
Kabul: 2023-10-17

Önerilen Atıf

Demir, E. K. (2023). Madde Sıralamasının bir öz yeterlik algısı ölçeğinin psikometrik özelliklerine etkisinin incelenmesi, *E-Uluslararası Eğitim Araştırmaları Dergisi,* 14 (5), 478-493. DOI: https://doi.org/10.19160/e-ijer.1362442

Genişletilmiş Özet

Problem: Bireylerin test puanlarını etkileyebilecek çeşitli hata kaynakları bulunmaktadır. Hata kaynakları genellikle ölçme aracının kendisi, ölçmenin yapıldığı koşullar veya testi alan bireylerin özellikleri olarak kabaca sınıflandırılabilir. Bu değişkenlerden gelen hatalar nedeniyle elde edilen ölçme sonucu gerçek performans düzeyinden farklılaşır. Bireylerin sınava girdikleri andaki kaygı, üzüntü, mutluluk gibi zihinsel durumları, sosyal beğenilirlik gibi toplumsal etkiler, yorgunluk, uyku yoksunluğu gibi fizyolojik etkiler, test puanlarının gerçek performans puanlarından sapmasına neden olabilir. Test geliştiriciler ölçme aracının kendisinden kaynaklanan ve kontrol edilmesi daha mümkün olan hataları kontrol altına almaya çalışması ölçmeye karışabilecek hataları azaltmaya yardımcı olur. Bir ölçme aracında bireylere verilen uyarıcılar temelde test maddeleridir ve bu maddeler tek başına ya da diğer maddelerle etkileşim halinde bireylerin verdikleri tepki ile test puanlarını oluşturur. Maddelerin hedeflenen yaş grubuna uygunluğu, ifadelerin açık ve net olması, yanıt kategorisinin uygunluğu gibi kontrol edilmediğinde hataya neden olabilecek pek çok değişken bulunmaktadır (Cronbach, 1988, 2013; Haladayna ve Rodriguez, 2013; Osterlind, 1998). Bunlardan en önemlilerinden biri de bir testteki maddelerin sıralamasının tepki davranışına ve dolayısıyla elde edilen puanlara etkisinin olup olmadığı problemidir.

Madde sıralama etkisi bir testteki daha önceki maddelerin daha sonraki yanıtları etkileyebileceği fikrini ifade eder (Schuman ve Presser, 1981). Madde sıralamasının etkisini ile ilgilenen ilk çalışmalar, başlangıçta maddelerin sıralamasından ziyade bireylere sunulan bilgi, kelime vb. uyaranların tepki davranışları üzerindeki etkisine odaklanmıştır. Bu çalışmalardan ilki Lund'a (1925) tarafından yapılan ve birbiriyle çelişen iki bilginin bireylere sunulmasının onların tutumlarını nasıl etkilediğini inceleyen araştırmadır. Lund'un (1925) çalışması madde sıralamasının bireylerin tutumları üzerinde tartışmalı konularda tartışmasız olanlara göre daha büyük bir etkiye sahip olduğunu göstermiştir (Crano, 1977). Bu çalışmadan sonra bireyler hakkında bilgi toplarken öncelik-sonralık etkisi araştırılmaya başlanmıştır (Chen, 2010). Brenner (1964) çalışmasında testin güvenirliğini, güçlüğünü ve ayırt ediciliğini incelemek için madde sıralamasını değiştirerek bir başarı testi uygulamıştır. Maddeler kolaydan zora, zordan kolaya veya rastgele sıralandığında testin güçlüğü, ayırt ediciliği ve güvenirliği açısından istatistiksel olarak anlamlı bir fark olmadığını bulmuştur. Benzer şekilde Munz ve Smouse (1968), öğrencilerin başarı testi puanlarını kolaydan zora, zordan kolaya ve rastgele olmak üzere üç farklı madde güçlük düzeniyle gözlemlemiş ve madde güçlüğüne göre sıralamanın toplam test puanı üzerinde bir etkisi olmadığı sonucuna ulaşmıştır. Klosner ve Gellman (1973) öğrencilerin sınav performanslarını üç test türüyle karşılaştırmıştır. Testlerde madde sıralaması konulara göre, konu içinde kolaydan zora, konu arasında kolaydan zora doğru sıralanmıştır ve test puanları arasında anlamlı bir fark olmadığını belirtmiştir. Hambleton ve Traub (1974) ise lise öğrencilerinin matematik testi performanslarını maddeleri kolaydan zora ve zordan kolaya doğru sıralayarak incelemişler ve öğrencilerin kolaydan zora doğru sıralanan maddelerden daha yüksek puanlar aldıklarını bulmuşlardır. Perlini vd. (1988) sınavlardaki bağlamsal etkileri araştırmıştır. Sınav süresi, madde sıralaması, konu sıralaması ve güçlük seviyesi gibi değişkenlerin test performansını nasıl etkilediğini incelemişlerdir. Sonuçlar, sınav süresinin, genel performans üzerinde etkisi olduğunu göstermiş ancak maddelerin konu veya güçlüğe göre sıralanmasının genel performans üzerinde etkisi olmadığını göstermiştir. Demirkol ve Kelecioğlu (2022), PISA 2015 Türkiye örnekleminde okuma ve matematik alanlarındaki madde konumunun etkisini araştırmışlardır. Sonuçlar, bu etkinin özellikle okuma alanında doğru cevap verme olasılığının azalmasına yol açtığını göstermektedir. Ayrıca matematik alanında açık uçlu maddelerde çoktan seçmeli maddelere kıyasla madde sıralama etkisinin daha fazla olduğu bulunmuştur.

Strack vd. (1988) tutum ölçeklerindeki madde sıralamasının büyük önem taşıyabileceğini vurgulamıştır. Knowles (1988) kişilik testlerinde madde sıralamasının etkisini araştırmıştır ve her bir maddenin mümkün olan her konumda sunulmasına olanak tanıyan birden fazla ölçek formu oluşturmuştur. Bulgular ortalama puanın madde sırasından etkilenmediğini göstermiştir. Vega ve O'Leary (2006), 641 öğrenciyle partner saldırganlığını incelemek amacıyla yürüttükleri araştırmada testin madde sırasını değiştirdikleri iki formunu uygulamışlardır. Test sonuçlarının madde sıralamasından etkilenmediği sonucuna varmışlardır. Chen (2010), tutum ölçeklerinde madde sıralamasının madde ayırt ediciliği, test puanları, test uzunluğu, yanıt süresi ve test güvenirliği üzerindeki etkisini araştırmak amacıyla bir çalışma yürütmüştür. Madde sıralamasının tutum ölçekleri üzerinde etkili olduğu, katılımcıların sonraki maddelere verdikleri yanıtların daha önce verilen maddelere bağlı olarak değişebileceğini ifade etmiştir.

Şahin (2021) araştırmasında siber aylaklık ölçeğini kullanarak aynı boyuta ait tüm maddelerin bir arada sunulduğu sabit sıralı bir form ve maddelerin rastgele sıralandığı bir form sunarak madde sıralama etkisini incelemiştir. Örneklemden alınan iki ayrı gruptan elde edilen sonuçlar, iki formun ortalama puanları arasında istatistiksel olarak anlamlı fark olduğunu ortaya koymuştur. Ölçeğin farklı madde sıralamasındaki formlarına ait ölçme değişmezliğinin sağlanıp sağlanmadığını incelenmiş ve madde sıralamasının farklı olduğu formlar arasında ölçme değişmezliğinin sağlanamadığı belirtilmiştir.

Madde sıralama etkisine yönelik çalışmalar incelendiğinde iki önemli problem ortaya çıkmaktadır. Öncelikle çalışmaların çoğunluğu maksimum performans testlerinde madde sıralama etkisine odaklanmış ve genellikle madde güçlüğü ile ilişkilendirilmiştir. Tutumlar, ilgi, yeterlik gibi tipik performansı hedefleyen ölçmeler üzerinde madde sıralama etkisi ile ilgili araştırmalar oldukça sınırlıdır ve çalışmalardan elde edilen sonuçlar birbiri ile çelişkilidir. Literatürde madde sıralama etkilerine ilişkin net bir cevap bulmak halen zordur ve bu durum hem psikolojik özellikleri ölçen ölçme araçlarının hem de özellikle eğitimde kullanılan başarı testlerinin geliştirilmesi ve uygulanması açısından önemli bir problemdir. Bu nedenle, bu çalışmanın amacı, tipik performansı ölçen bir öz yeterlik algısı ölçeği kullanarak madde sıralama etkilerini araştırmaktır. Bir diğer problem ise madde sıralama etkisi ile ilgili yapılan çalışmalarda aynı gruptan veri toplayan çalışmaların çok sınırlı olması ve çalışmaların çoğunun üniversite öğrencileriyle yapılmış olmasıdır. Dolayısıyla bu problemin çözümüne katkı sağlamak amacıyla bu çalışmanın katılımcıları ortaokul 8. sınıf öğrencileri olup, aynı grup üzerinde madde sıralama etkisi incelenmiştir. Ölçeğin farklı madde sıralamasına sahip iki formu arasında ölçeğin psikometrik özelliklerinin korunup korunamadığı, grubun ortalama puanlarında anlamlı bir değişik olup olmadığı ve farklı madde sırasına sahip formlar arasında ölçme değişmezliğinin sağlanıp sağlanamadığı ile ilgili sorulara yanıt aranmıştır.

Yöntem: Bu araştırmada Ortaokul 8. sınıf öğrencileriyle gerçekleştirilmiş olup iki veri toplama aşamasından oluşmaktadır. İlk aşamada çalışmaya 1088 öğrenci katılmıştır. Ancak 64 öğrenci her iki uygulamaya birden katılamadığı için, 78 öğrenci ise maddelerin yarısından fazlasına cevap vermediği için çalışmaya dahil edilmemiştir. Sonuç olarak bu araştırma %51'i kadın, %49'u erkek olmak üzere 946 öğrenci ile gerçekleştirilmiştir. Çalışmada Gülbahar vd. (2019) tarafından ortaokul öğrencileri için geliştirilen Bilgi İşlemsel Düşünme Becerisine Yönelik Özyeterlik Algısı (BİDBÖA) Ölçeği kullanılmıştır. Ölçeğin beş alt boyutu bulunmaktadır ve toplamda üç dereceli yanıt kategorisine sahip 36 maddeden oluşmaktadır. Ölçeğin alt boyutları sırasıyla Algoritma Tasarlama Yeterliği (9 madde), Problem Çözme Yeterliği (10 madde), Veri İşleme Yeterliği (7 madde), Temel Programlama Yeterliği (5 madde) ve Özgüven Yeterliği (5 madde) olarak isimlendirilmiştir. Ölçeğin orijinal formunda bu beş boyutun her birine karşılık gelen maddeler, ilgili boyut altında toplu olarak sunulmaktadır. Madde sıralama etkisini ölçmek için maddeler rastgele yeniden düzenlenmiş ve rastgele-form olarak adlandırılan başka bir form oluşturulmuştur. Başlangıçta tüm öğrencilere orijinal-sıralı form uygulanmış, ardından bir ay sonra ikinci oturumda rastgele-sıralı form uygulanmıştır. Araştırmaya başlamadan önce ölçeği geliştiren yazarlardan izin alınmıştır. Ölçeğin uygulama aşamasında öğrencilerden onam formları alınmış ve çalışmaya sadece gönüllü öğrenciler katılmıştır. İsim bilgileri de dahil olmak üzere hiçbir kişisel veya demografik bilgi toplanmamıştır. Katılımcılar formları bir rumuz seçerek doldurmuşlar ve her iki ayrı başvurudaki formlar da bu rumuz kullanılarak eşleştirilmiştir.

Araştırmanın analizleri için, öğrencilerin farklı madde sıralamasına sahip formlardan elde ettikleri puan ortalamaları arasında anlamlı bir fark olup olmadığını test etmek amacıyla eşleştirilmiş örneklem t-testi kullanılmıştır ve bu puanlar arasındaki korelasyon incelenmiştir. Her bir form için ölçeklerin yapısı, faktör yükleri ve model uyum indeksleri incelenerek doğrulayıcı faktör analizi (DFA) yapılmıştır. Ayrıca bu çalışmada ölçeğin güvenirliğini incelemek için McDonald's ω, Cronbach α ve Tabakalı α katsayıları hesaplanmıştır. Ölçek formları arasındaki ölçme değişmezliğini incelemek için ise çok gruplu DFA (çg-DFA) analizi yapılmıştır. Ölçme değişmezliğini değerlendirmek için ki-kare fark testleri ve modeller arasındaki CFI, RMSEA fark analizleri kullanılmıştır. Ki-kare değişim değerinin $p>.05$ ve CFI, RMSEA ≤ 0.01 olması, o model için ilgili düzeyde ölçme değişmezliğinin sağlandığını göstermektedir (Cheung ve Rensvold, 2002; Hu ve Bentler, 1998). Analizler Mplus 8.0 (Muthén ve Muthén, 2017) ile yapılmıştır.

Bulgular: Orijinal-sıralı form ile rastgele-sıralı form toplam puanları arasında çok yüksek pozitif korelasyon (r=0.98) bulunmaktadır. Benzer şekilde alt boyutlar açısından da her iki form arasındaki

korelasyon katsayısı 0.79 ile 0.72 arasında değişmektedir. Toplam puanlar arasında pozitif korelasyonun yüksek olması, puanlar arasında anlamlı bir fark olmadığı anlamına gelmediğinden, BİDBÖA ölçeğinin hem orijinal hem de rastgele sıralı formlarından elde edilen puanlara ilişkin t-testi analizi sonucunda orijinal sıralama formunu alan katılımcıların toplam ölçek ortalama puanları (M = 80.95, SD = 14.27) ile rastgele sıralama formunu alanlar (M = 80.83, SD = 14.35) arasında istatistiksel olarak anlamlı bir fark olmadığı t(945)= 1.67, p=0.10 görülmüştür. Benzer şekilde hiçbir alt boyut puanları arasında da anlamlı fark görülmemiştir (p>.05). Her bir öğrenci için her iki formdan aldığı puanlar arasındaki fark ile bu farkların ortalaması hesaplanmıştır. Alt boyutlar için ortalama fark puanları sırasıyla 2.25, 2.69, 2.35, 1.98 ve 2.05 olarak hesaplanmıştır. Her öğrencinin iki formdan aldığı toplam puanlar arasındaki farkın ortalaması ise 0.89 olarak bulunmuştur. İç tutarlık anlamında güvenirliği incelemek amacıyla her iki form için Cronbach α, McDonald's ω ve Tabakalı α katsayıları hesaplanmıştır. Ölçeğin tamamı için tabakalı α katsayısı kullanılmış olup orijinal-sıralı formu için 0.94, rastgele-sıralı formu için 0.93 olarak hesaplanmıştır. Alt boyutlar için hesaplanan Cronbach α katsayıları orijinal-sıralı form için 0.92 ile 0.75 arasında değişirken, rastgele-sıralı form için 0.90 ile 0.79 arasında değişmektedir. Her iki form için de ölçekten alınan puanların iç tutarlık anlamında güvenilir olduğu ifade edilebilir. Ancak formlar karşılaştırıldığında tabakalı α katsayısının orijinal sıralı form için daha yüksek olduğu benzer şekilde tüm alt boyutlar için de Cronbach α katsayısının orijinal-sıralı formda daha yüksek olduğu ifade edilebilir.

Ölçek formları DFA ile incelendiğinde her iki form için de model uyumunun sağlandığı ifade edilebilir. CFI ve TLI değerleri açısından her iki formun model uyum değerleri aynıdır. Ancak χ2/df oranı, SRMR ve RMSEA değerleri dikkate alındığında orijinal-sıralı formun model uyumu rastgele-sıralı forma göre daha iyidir. Her iki modele ait DFA sonucunda elde edilen madde faktör yükleri incelendiğinde ilk dört alt boyutta faktör yüklerinin şaşırtıcı derecede benzer olduğu görülmektedir ve faktör yükleri 0.81 ile 0.51 arasında değişmektedir. Özgüven Yeterliği alt boyutunda ise orijinal-sıralı formundaki tüm madde faktör yükleri diğer forma göre daha yüksektir. Tüm faktör yükleri arasında en düşük değer ise 0.47 değeriyle rastgele-sıralı formdadır. Aynı maddenin orijinal-sıralı formundaki faktör yükü 0.53 olarak hesaplanmıştır. Ölçme değişmezliğini incelemek çg-DFA ile ki-kare, CFI ve RMSEA fark testleri kullanılmıştır. Ki-kare değişim değerinin p>.05 ve CFI, RMSEA ≤ 0.01 olması ölçme değişmezliğinin ilgili düzeyde sağlandığını göstermektedir (Cheung ve Rensvold, 2002; Hu ve Bentler, 1998; Putnick ve Bornstein, 2016). Ki-kare fark testinin tüm anlamlılık değerleri p>.05 olarak hesaplanmıştır. Aynı zamanda tüm ölçme değişmezliği seviyeleri için ΔCFI ve ΔRMSEA değerleri 0.01'den küçüktür. Bu durumda orijinal-sıralı ve rastgele-sıralı formlar arasında katı düzeyde ölçme değişmezliğinin olduğu ifade edilebilir.

Sonuç, Tartışma ve Öneriler: Bu çalışmada, bilgi işlemsel düşünme becerisini öz yeterlilik algısını ölçen bir öz yeterlik algısı ölçeği olan BİDBÖA ölçeğinin aynı gruba önce orijinal sırasıyla, daha sonra tamamen rastgele sıralı formun uygulanmasıyla madde sıralama etkisi araştırılmıştır. İki form arasında elde edilen puanlar arasındaki fark incelenmiş, toplam puanlar arasında anlamlı bir fark bulunamamıştır. Bu sonuç; Munz ve Smouse (1968), Krampen vd. (1992), O'Shaughnessy (1983), Laffitte (1984), Solomon ve Kopelman (1984), Vega ve O'Leary (2006), Etzel vd. (2021), Hartig vd. (2007), Albano (2013), Asseburg ve Frey'in (2013) bulguları ile örtüşmektedir ancak Bossart ve Di Vesta (1966), Flaugher vd. (1968), Hambleton ve Traub (1974), Strack vd. (1988), Dillman (2000) ve Şahin'in (2021) bulguları ile farklılık göstermektedir. Grup ortalamaları arasında fark olmaması ve bireysel puan farkı ortalamaları arasındaki farkın az olması nedeniyle madde sırasının puanlar üzerinde anlamlı bir etkisinin olmadığı söylenebilir. Ancak bu çalışmanın öz yeterlik algısı ölçeği ile yürütüldüğünü, puan farklarının çok önemli olduğu ölçme durumlarını temsil etmediğini de göz önünde bulundurmak gerekir. Ölçeklerin güvenirlik düzeyleri incelendiğinde ölçeğin tamamı için tabakalanmış α değerinin orijinal-sıralı formda biraz daha yüksek olduğu ancak her iki form için de oldukça yüksek katsayılar elde edildiği ifade edilebilir. Alt boyutlara ait Cronbach α katsayıları incelendiğinde, orijinal sıralı form değerlerinin tüm alt boyutlar için daha yüksek olduğu görülmüştür. McDonald's ω'da ise ilk dört alt boyuta ait değerler aynı iken son alt boyuta ait orijinal sıralama formu değeri daha yüksektir. Ölçme aracında aynı boyutu temsil eden maddelerin ardı ardına verilmesi iç tutarlık güvenirliğine katkı sağlamıştır. Benzer şekilde Şahin (2021) siber aylaklık ölçeğini kullanarak yaptığı araştırmada sabit-sıralı formda daha yüksek iç tutarlık katsayıları bulmuştur. Madde sıralamasının ölçeğin yapısına ve faktör yüklerine etkisini incelemek amacıyla yapılan DFA ile her iki forma ait modelin de benzer model uyum indeksi değerleri ile iyi bir model uyumu göstermesi, maddelerin rastgele seçilmesinin ölçeğin beş alt boyuttan oluşan yapısını bozmadığını göstermiş ve aynı

yapı korunmuştur. Faktör yükleri incelendiğinde ilk dört alt boyut için faktör yüklerinin birbirine oldukça yakın olduğu görülmüştür. Özgüven Yeterliği adı verilen son alt boyutta ise orijinal-sıralı formun faktör yükleri daha yüksektir. Aynı faktörü temsil eden maddelerin bir arada bulunmasının sadece bu alt boyut için faktör yüklerine olumlu etki yaptığı söylenebilir. Formlar arasındaki ölçme değişmezliğini test etmek için yapılan çg-DFA analizi sonuçları, iki model arasındaki ölçme değişmezliğinin pratikte elde edilmesi zor bir düzey olan katı ölçme değişmezliği düzeyinde elde edildiğini göstermektedir. Dolayısıyla BİDBÖA ölçeğinde maddelerin sıralı ya da rastgele verilmesi arasında ölçme değişmezliğinin olduğu ifade edilebilir.

Literatürde madde sıralama etkisi ile ilgili yapılan çalışmalarda farklı ve çelişkili sonuçların elde edildiği görülmektedir. Bazı çalışmaların madde sıralama etkisinin ortalama puan değişimi, ölçek yapısının korunması, iç tutarlık, ölçme değişmezliği gibi psikometrik özellikler ile ilgili tutarlı sonuçlar verdiği, bazılarının ise çelişkili sonuçlar ürettiği görülmektedir. Bu durum çalışmalarda farklı ölçme araçlarının kullanılmasından kaynaklanabilir. Maksimum performansın ölçüldüğü çalışmalarda maddelerin zorluk derecesi daha önemli hale gelirken, tipik performansı ölçen testlerde maddeler arasındaki bağlamsal etkiler daha fazla önem kazanmaktadır. Madde sıralama etkisi bu iki test türünde farklı sonuçlar doğurabilmektedir. Tutum, ilgi, yeterlik gibi birbirinden farklı özellikleri ölçen tipik performans ölçeklerinde de farklı madde sıralama etkileri görülebilmektedir. Bu çalışmada bilgi işlemsel düşünme becerisi öz yeterlik algısına ilişkin bir ölçme aracı kullanılmıştır. Öz yeterlik algısı madde sırasının etkisinden daha az etkilenen bir özellik olabilir. Bu nedenle madde sıralama etkisinin önemi ölçülen özelliğe bağlı olarak değişebilir. Gelecek araştırmalarda özellikle tipik performansın ölçülmesinde ölçülen özelliğin farklılaşmasına odaklanan çalışmalar yapılabilir. Tutum ölçeklerinde maddelerin bağlamsal etkilerini araştırmak için sıralamanın rastgele verilmek yerine belirli bir amaç doğrultusunda madde sırasının manipüle edildiği farklı formlar kullanılabilir. Örneğin, olumsuz maddeler bir araya toplanarak madde yanıtları üzerindeki etkileri incelenebilir. Bu çalışmada orijinal-sıralı ve rastgele-sıralı formlar aynı gruba belirli bir zaman aralığıyla uygulanmıştır, ileride yapılacak araştırmalarda ön test etkisini önlemek amacıyla formlar iki ayrı grupta çapraz olarak uygulanabilir. Ayrıca madde sıralama etkisi farklı örneklem büyüklüklerinde, yaş gruplarında, demografik özelliklerde ve kültürlerde incelenebilir. Madde türü ile madde sıralama etkisi arasındaki ilişkiyi ortaya çıkarabilecek araştırmalar da alan yazına ve test geliştiricilere önemli katkı sağlayacaktır.

KAYNAKÇA/REFERENCES

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*(4), 408-426. https://doi.org/10.1111/jedm.12026

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*(1), 92.

Bornstein, M. H. (1995). Form and function: Implications for studies of culture and human development. *Culture & Psychology*, *1*(1), 123-137. https://doi.org/10.1177/1354067X95110

Bossart, P., & Di Vesta, F. J. (1966). Effects of context, frequency, and order of presentation of evaluative assertions on impression formation. *Journal of Personality and Social Psychology*, 4, 538-544. https://doi.org/10.1037/h0023898

Bradburn, N. M. (1983). Response effects. *Handbook of survey research*, 1, 289-328.

Brenner, M., H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, 48, 98-100. https://doi.org/10.1037/h0045738

Chen, P. H. (2010). Item order effects on attitude measures (Doctoral dissertation, University of Denver).

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Crano, W. D. (1977). Primacy verusus recency in retention of information and opinion change. *Journal of Social Psychology*, 101, 87-96. https://doi.org/10.1080/00224545.1977.9923987

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika*, 53, 63-70.

Cronbach, L. J. (2013). Five perspectives on validity argument. In *Test Validity* (pp. 3-17). Routledge.

Demirkol, S., & Kelecioğlu, H. (2022). Investigating the effect of item position on person and item parameters: PISA 2015 Turkey sample. *Journal of Measurement and Evaluation in Education and Psychology*, 13(1), 69-85.

https://doi.org/10.21031/epod.958576

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method*. New York: John Wiley & Sons, Inc.

Etzel, J. M., Holland, J., & Nagy, G. (2021). The internal and external validity of the latent vocational interest circumplex: Structure, relationships with self-concepts, and robustness against item-order effects. *Journal of Vocational Behavior*, 124, 103520. https://doi.org/10.1016/j.jvb.2020.103520

Flaugher, R., L., Melton, R., S., & Myers, C., T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement*, 28, 813-824. https://doi.org/10.1177/001316446802800310

Fraenkel, J. R.,& Wallen, N. E. (2009). *How to design and evaluate research in education.* San Fransisco: Mc-Graw Hill Pub.

Glanville, J. L., & Wildhagen, T. (2007). The measurement of school engagement: Assessing dimensionality and measurement invariance across race and ethnicity. *Educational and Psychological Measurement*, 67(6), 1019-1041. https://doi.org/10.1177/00131644062991

Gülbahar, Y., Kert, S. B., & Kalelioğlu, F. (2019). Bilgi işlemsel düşünme becerisine yönelik öz yeterlik algısı ölçeği: Geçerlik ve güvenirlik çalışması. *Turkish Journal of Computer and Mathematics Education* (TURCOMAT), 10(1), 1-29. https://doi.org/10.16949/turkbilmat.385097

Haladyna T. M., Rodriguez M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *Journal of Experimental Education*, 43, 40-46. https://doi.org/10.1080/00220973.1974.10806302

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research*, 42(1), 157-183. https://doi.org/10.1080/00273170701341266

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3), 117-144. https://doi.org/10.1080/03610739208253916

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under parameterized model misspecification. *Psychological Methods*, 3(4), 424. https://doi.org/10.1037/1082-989X.3.4.424

Jöreskog, K. G., & Sörbom, D. (1993*). LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific software international. Chicago, IL, US.

Klosner, N. C., & Gellman, E. K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement*, 33, 413-418. https://doi.org/10.1177/001316447303300022

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, 55, 312-320. https://doi.org/10.1037/0022-3514.55.2.312

Krampen, G., Hense, H., & Schneider, J. F. (1992). Reliabilität und Validität von Fragebogenskalen bei Standardreihenfolge versus inhaltshomogener Blockbildung ihrer Items. Zeitschrift für experimentelle und angewandte Psychologie.

Laffitte Jr., R. G. (1984). Effects on item order on achievement test scores and students' perception of test difficulty. *Teaching of Psychology*, 11, 212-214. https://doi.org/10.1177/0098628384011004

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., ... & van de Schoot, R. (2022). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. Social Science Research, 102805.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.

Lund, F. H. (1925). The psychology of belief: A study of its emotional, and volitional determinants. *The Journal of Abnormal and Social Psychology*, *20*(2), 174. https://doi.org/10.1037/h0066996

Luong, R., & Flake, J. K. (2022). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods.* 28(4), 905–924. https://doi.org/10.1037/met0000441

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses*. Psychological Methods,* 7(1), 64. https://doi.org/10.1037/1082-989X.7.1.64

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.

Munz, D. C., & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance*. Journal of Educational Psychology*, 59, 370-374. https://doi.org/10.1037/h0026224

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide*. 8th Edn. Los Angeles, CA: Muthén & Muthén

Osterlind, S. J. (1998). *What is constructing test items?* (pp. 1-16). Springer Netherlands.

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie Canadienne*, 39(4), 299. https://doi.org/10.1037/h0086821

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Schuman, H., & Presser, S. (1981). The attitude-action connection and the issue of gun control. *The Annals of the American Academy of Political and Social Science*, 455(1), 40-47. https://doi.org/10.1177/000271628145500105

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two dimensional model of APM. *Psychology Science Quarterly,* 51(1), 47–64.

O'Shaughnessy, E. (1983). Words and working through. *The International Journal of Psycho-Analysis*, 64, 281.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, 18(5), 429-442. https://doi.org/10.1002/ejsp.2420180505

Solomon, E., & Kopelman, R. E. (1984). Questionnaire format and scale reliability: An examination of three modes of item presentation. *Psychological Reports*, 54(2), 447-452. https://doi.org/10.2466/pr0.1984.54.2.447

Şahin, M. D. (2021). Effect of item order on certain psychometric properties: A demonstration on a cyberloafing scale. *Frontiers in Psychology*, 12, 590545. https://doi.org/10.3389/fpsyg.2021.590545

Vega, E. M., & O'Leary, K. D. (2006). Reaction time and item presentation factors in the self-report of partner aggression. *Violence and Victims*, 21, 519-532. https://doi.org/10.1891/vivi.21.4.519

Weinberg, M. K., Seton, C., & Cameron, N. (2018). The measurement of subjective wellbeing: Item-order effects in the personal wellbeing index—adult. *Journal of Happiness Studies*, 19, 315-332. https://doi.org/10.1007/s10902-016-9822-1

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, 4(1), 10-18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009