International Journal of Contemporary Educational Research (IJCER)

www.ijcer.net

Development of Test-Taking Strategies Scale: High School and Undergraduate Form

Emine Burcu Tunç [1], Selma Şenel [2]
[1]Marmara University, 0000-0002-8225-9299
[2] Balıkesir University, 0000-0002-5803-0793

Article History
Received: 28.02.2021
Received in revised form: 20.09.2021
Accepted: 30.09.2021
Article Type: Research Article

To cite this article:

Tunç, E. B. & Şenel, S. (2021). Development of Test-Taking Strategies Scale: High School and Undergraduate Form. *International Journal of Contemporary Educational Research*, *8*(4), 116-129.https://doi.org/10.33200/ijcer.888368

International Journal of Contemporary Educational Research
Volume 8, Number 4, December 2021, Page 116-129 ISSN: 2148-3868

Development of Test-Taking Strategies Scale: High School and Undergraduate Form

Emine Burcu Tunç[1*], Selma Şenel[2]
[1]Marmara University
[2]Balıkesir University

Abstract

Test-taking strategies are discussed in the literature as an important factor affecting test scores and are recommended to be taken into consideration regarding the validity of tests. Although studies have been conducted for more than a quarter century, no agreement has been reached on the dimensions of test-taking strategies. The purpose of this study is to develop a valid and reliable scale of test-taking strategies for university and high school students who experience intense periods of testing. In the scale created for this purpose, we consider tests with different types of titems and focus on strategies before, during, and after the test, excluding test preparation. Two separate forms of test-taking strategies were developed for the high school (27 items) and undergraduate (18 items) levels, using promising measurement theories and models.. Results indicated that the Person Separation Index, as a reliability index, was calculated as .88 for the high-school form and .93 for the undergraduate form. This study is significant in presenting valid and reliable tools for measuring test-taking strategies and can be considered exemplary research that uses the Partial Credit Model for Likert-type scale development.

Keywords: Test-taking strategies, Scale development, Likert type, Item Response Theory, Partial Credit Model

Introduction

Students face numerous tests and examinations throughout their educational life. Test scores are used for a variety of purposes, such as determining course performance, certification, admission to higher levels of education such as a university, or when seeking employment. Especially, the results of high-stakes tests such as a university entrance exam or civil service personnel selection exam, in the case of Turkey, are of significant importance in terms of their influence and impact on students' future lives.

The main objective of educational testing is to measure students' competency related to certain traits measured by the test. In general, students who demonstrate boundless efforts achieve the required competency and obtain significantly high scores from exams. It has been significantly reported that students may spend years preparing for high-stakes tests, with content-focused publications generated for full test preparation (Educational Testing Service, 2001). Practice, as part of exam preparation can have an important effect that reinforces learning. However, apart from the measured traits, numerous cognitive, psychological, physiological, and environmental variables may affect test results such as motivation, self-efficacy, perception, test anxiety, physical disability, or test-taking strategies. There are also numerous variables outside the focus of an assessment that can affect the results of exams. Test validity is theoretically defined as 'the characteristics of the examiner that are not related to measured trait' that may affect the test results. For this reason, creating standard test conditions to increase the validity of measurement results (American Psychological Association, American Educational Research Education, National Council on Measurement in Standards, 1966), providing test adaptations for people with disabilities (Şenel & Kutlu, 2018; Şenel & Şenel, 2018), using test anxiety counselling programmes (Demirci & Erden, 2016), and teaching test-taking strategies (Beidel et al., 1999; Dodeen, 2015; Kesselman-Turkel & Peterson, 2004) are becoming increasingly important. Therefore, it is important to be able to measure such constructs that may affect test scores and provide evidence of a test's validity.

On the validity of test scores, the literature focuses on the content of the measurement, test plan preparation, test development, and test statistics. However, respondents' answering behaviours and test-taking strategies are not considered within the focus of research and discussions regarding test validity (Bachman, 1990; Cohen, 2007).

* Corresponding Author: *Emine Burcu Tunç, burcupehlivantunc@gmail.com*

Where some respondents effectively employ appropriate test-taking strategies, other respondents at the same proficiency level may face difficulties regarding what the exam measures, in other words, the validity of the test. Test-taking strategies have been discussed in the literature as an important factor affecting test scores (Beidel et al., 1999; Dodeen et al., 2014; Therrien et al., 2009), and are recommended to be taken into consideration in terms of the validity of tests (Bachman, 1990; Cohen, 2007). Test-taking strategies can also be evaluated as a part of the test itself (Cohen, 2007).

It should be considered how strategic approaches may be used for different test item types, which may affect the score and validity of the exam. 'Strategy' can commonly be defined as a set of tactics and methods applied to achieve a specified goal (Ün Açıkgöz, 2003). On the other hand, *test-taking strategies* consist of various information, techniques, and methods used to answer test items, apart from the cognitive skills of the respondent, to achieve exam success or to gain a higher test score. In the literature, such strategies are referred to as *test-taking strategies* (Chittooran & Miles, 2001; Dodeen, 2015; Hong et al., 2006; Peng et al., 2014; Therrien et al., 2009) or *test-wiseness strategies* (Cohen, 2007), and as a more inclusive term *test-taking skills* (Boyd, 1989; Chittooran & Miles, 2001; Dodeen et al., 2014; Lewandowski et al., 2013). In the current study, it was found appropriate to apply the term *test-taking strategies*. However, the scale's theoretical framework includes all testing types as simply the 'exam', with *exam,* considered a much broader concept than simply *testing*. While 'test' is used for a specific measurement tool or technique, 'exam' refers to the entire end-to-end assessment process. For example, university entrance exams commonly consist of quantitative and verbal tests. *Exam* is a more inclusive expression and is used as a term that explains the application of tests and the entire assessment process (Tekin, 1996). In contrast, *test* is a more general label that covers measurement tools much broader (Baykul, 2010).

The literature explains the relationship between test-taking strategies and various psychological characteristics, and these studies can be summarised as follows:

- There is a negative correlation between test-taking strategies and exam anxiety (Bruch, 1981; Dodeen et al., 2014; Dodeen, 2015; Peng et al., 2014). Exam anxiety can be reduced through training on test taking strategies (Beidel et al., 1999; Chittooran & Miles, 2001; Dodeen, 2015; Lewandowski et al., 2013). Kesselman-Turkel and Peterson (2004) and Chittooran and Miles (2001) also considered the reduction of test anxiety as a form of test-taking strategy.
- Using test-taking strategies has increased exam scores (Bruch, 1981; Lewandowski et al., 2013; Therrien et al., 2009) and positive attitudes towards exam-taking (Dodeen, 2015).
- Low-achiever students tend to use test-taking strategies more (Cohen, 2007).
- It is important to teach test-taking strategies to students with special needs (Lewandowski et al., 2013; Therrien et al., 2009), otherwise such strategies are unlikely to be adopted. Using test-taking strategies can prevent students with special needs from falling behind their peers due to a lack of strategies.The use of test-taking strategies indicates a positive correlation with course motivation (Peng et al., 2014).
- Recent studies on test-taking strategies focus on technologies that enable individuals to record variables such as time spent answering and eye movements in computer-based tests (Brunfaut & McCray, 2015; Roderer & Roebers, 2014).

The first step in a scale development study is to define the construct to be measured and establish its theoretical basis (Erkuş, 2012). As observed in the literature, there are various conceptual definitions of test-taking strategies and theoretical models based on different dimensions. However, no consensus has been reached on the dimensions of test-taking strategies, although this area has been studied for more than a quarter of a century (Cohen, 2007). The dimensions of studies that have looked at test strategies over the past 20 years and the instruments and techniques they have used to determine test strategies are summarised in Table 1.

When Table 1 and similar studies in the literature (e.g., Kesselman-Turkel & Peterson, 2004) are examined, the major test-taking strategy dimensions can be summarised in three different approaches. First, some deal separately with affective, cognitive, and metacognitive dimensions. The importance attributed to the exam, self-efficacy, test motivation, and attitude are the affective characteristics that are particularly emphasised for reducing test anxiety. The cognitive dimension refers to the cognitive processes employed whilst answering, other than the measured feature. Metacognitive strategies can be expressed as the ability to be aware of the students' cognitive and affective strategies, organise them. Many studies are based solely on the metacognitive dimension. Second, some studies consider the test preparation as test-taking strategies; however, since test preparation includes a dimension that could also be considered as studying skills, it may not be possible to focus solely on the exam. It should also be taken into account that some students follow the lessons and can make an exam-oriented preparation by carrying out a planned study (Yıldırım et al., 2000). Apart from this, classifying test-taking strategies as pre-test, during-test and post-test is another accepted classification (Dodeen, 2008).

Table 1. Summary of the research focusing on test strategies

Research	Peng et al. (2014)	Bıçak (2013)	Hong et al. (2006)	Dodeen (2008)	Chittooran & Miles (2001)
Dimensions	*Motivational* • Importance of exam • Effort • Self-efficacy • Test anxiety *Cognitive* • Tactics • Metacognitive strategies	*Test preparation* • Cognitive • Social • Metacognitive *During test* • Item analysis • Time scheduling • Correct response estimation *Post-test*	• Test preparation strategies • Test preparation awareness • Test-taking strategies	• Pre-test strategies • Strategies during testing • Post-test strategies • Time management	• Familiarity with test features • Familiarity with test content • Test preparation • Test wisdom • Management of test anxiety
Data Collection Tool	*Test-Taking Strategies Questionnaire* (Hong & Peng, 2004) and applied by adding new items for research purposes	• *Test Preparation Scale* • *Test-Taking Strategies Scale* (secondary education students)	Interview	31-item scale (university students)	Literature review

Studies aimed at determining test-taking strategies are mostly conducted with qualitative research or using checklists and questionnaires based upon self-evaluation and perception (Cohen, 2007; Hong et al., 2006; Pehlivan & Kutlu, 2014; Peng et al., 2014). More recent research studies have assessed test-taking strategies using psychological measurement tools (Bıçak, 2013; Dodeen, 2008). As test-taking strategy vary according to the item types of the exams (Anderson, 1991; Boyd, 1989; Cohen, 2007; Kesselman-Turkel & Peterson, 2004), scale items differentiate according to exam item types. For instance, a strategy for multiple-choice items such as 'When answering questions, I eliminate the option that looks different from the others' cannot be applied in an exam consisting of open-ended items.

The literature includes scale development studies to determine test preparation strategies and test strategies, and various studies that have applied the developed scales (e.g., Bıçak, 2013; Dodeen, 2008; Dodeen et al., 2014). However, the focus of these studies was on high school (Bıçak, 2013) and university students (Dodeen, 2008), and the literature has mostly discussed strategies for multiple-choice items excluding open-ended items and other item types (Bıçak, 2013). Furthermore, the studies in which test preparation strategies were examined were mainly based on indicators of studying skills. Thus, the goal of the current research is to develop scales that include test-taking strategies for pre-test, during-test, and post-test, but not test preparation strategies for high difficulty exams that university and high school students often face and that have different item types.

In order to collect validity and reliability evidence of the scale development process, techniques based mainly on the Classical Test Theory are used. For the reliability proofs of a Likert-type scale, the Cronbach alpha or split-half methods are mainly used as an indicator of internal consistency, and item-total correlations are presented as a statistical value for item validity (Kartal & Dirlik, 2016; Kizilkaya & Aşkar, 2009; Kutlu et al., 2009). Classical test theory has its limitations as it provides values that depend on the study group or item sample, focuses on only one source of error (the internal consistency indicator Cronbach's alpha focuses on the consistency of item scores), and provides a single reliability value (Crocker & Algina, 2006; Embretson & Reise, 2000; Hambleton et al., 1991; Kaya Uyanık et al., 2019). The Item Response Theory (IRT), which largely exceeds these limitations, is a powerful theory widely used among current measurement theories.

In light of the latest developments in measurement and evaluation, although there has been a slight increase in the use of IRT-based models in the development of Likert-type scales, there has been limited research undertaken in this area (Demirtaşli et al., 2016; İlhan & Güler, 2018; Wongpakaran et al., 2020; Yaşar & Aybek, 2019). There are many advantages suggested for the use of the Rasch model in the process of collecting validity evidence for a Likert-type scale (Bond & Fox, 2015; Boone et al., 2014; Engelhard & Wind, 2017; Güler, 2014; İlhan & Güler, 2018; Linacre, 1994; Primi et al., 2019).

In the current study, Partial Credit Model (PCM) was used, one of the models based on IRT. PCM has both the advantages of IRT and the features of the Rasch model. It was developed by Masters in 1982, and is an extension of the Rasch model developed for two-category items. This model is used when distances between the response categories in Likert-type items differ from item to item. One of the important features of the model is that it is possible to score individuals with a moderate level θ (Koch & Dodd, 1989). The use of PCM is strongly recommended due to its advantages over IRT (Van Zile-Tamsen, 2017). In addition to the main purpose of the

research, this study aims to contribute to the literature by reflecting current and valid measurement approaches in the field and providing an example of Likert-type scale development based on the Partial Credit Model (PCM).

Method

In the current study, we aimed to develop a measurement tool to determine students' test-taking strategies. In this context, this research is a scale development study. Information about the study group and the processes followed throughout the development of the test-taking strategies scale are as follows.

Study Group

In scale development studies, the trial application group should be as heterogeneous as possible regarding the feature to be measured (Erkuş, 2012). In this way, statistical results can be examined for their ability to measure individuals who have the measured characteristic at different levels. For this reason, we choose a working group that would include individuals using different strategies at different levels. The scale was chosen to include the high school and undergraduate students of the group it was developed. It is thought that these groups may show different characteristics in being exposed to different types of exams and test-taking strategies. A total of 321 high school students in their final grade (i.e., 12^{th} grade) from Anatolian, Science, Social Sciences, and Vocational high schools in Turkey were reached with convenience sampling. 71% are female (n = 229) and 29% male (n = 92). Additionally, 337 undergraduate students attending Tourism, Education, Engineering and Science, and Literature faculties were reached, with 68% of the students being female (n = 231) and 32% male (n = 106). Additionally, 110 students-49 high school students for the high school form and 61 university students for the undergraduate form the study to examine the criterion-referenced validity of the final forms.

Development of the Items

A review of the different methods used to examine testing strategies in the literature can be found in Table 1. The items in this study targeted pretest, duringtest, and posttest strategies based on the scales, questionnaires, and findings used in the literature presented. The reason for developing items that take into account these three different time intervals is that strategies differ at certain points in the testing process. Prior to an exam, students may use certain strategies to prepare themselves physiologically, psychologically, and cognitively. These strategies include, for example, consuming drinks that they believe will increase their alertness, trying to relax by taking a walk in the fresh air, and discussing controversial topics with friends. During an exam, the primary goal is to answer as many questions as accurately and completely as possible. Following an exam, it is about evaluating the answers given and assessing the strategy used during the exam by monitoring one's time management, reviewing any mistakes, and organising or changing strategies before and during the exam to better prepare for the next exam.

While developing scale items, the literature (Chittooran & Miles, 2001; Cohen, 2007; Dodeen, 2015; Rozakis, 2003; Yıldırım et al., 2000) and items from similar scales in the literature (Bıçak, 2013; Dodeen, 2008) were used in the current study. The scale was developed as a 5-point, Likert type instrument consisting of *never*, *rarely*, *sometimes*, *often*, and *always* response categories. Following the item writing process, a 49-item trial form was created. The form was then reviewed by three lecturers from the field of Assessment and Evaluation, and one faculty member from the field of Guidance and Psychological Counselling, in terms of reflecting the relevant structure of the items, the accuracy of the statements used, and whether or not the scope was reflected adequately and accurately. Finally, as the scale was developed and applied in Turkish, the language and clarity were evaluated and edited by a faculty member from the Turkish Language and Literature department. With revisions taking into account the expert opinions received, the form was subsequently reduced to 47 items.

The trial form was then applied as an online instrument. A pre-trial application was first applied to a total of 19 students (eight high school and 11 undergraduates) to observe in advance any unforeseen issues with comprehensibility or implementation. The participants found the trial form to be mostly clear and understandable. However, one respondent stated having to read Item 23 several times to understand it. This item was subsequently changed to a more simplified structure. The original items (included and excluded) are presented in Appendix 5.

Data Collection

A trial application is the process of collecting data for validity proofs of the scale. In this process, participant volunteerism is very important as the accuracy of the data affects the structure of the final scale. Sending out the online form of the scale electronically and requiring no personal information may provide the necessary freedom

for volunteering; however, education level, faculty and department, gender, and grade level were obtained from the participants for analysis.

Data Analysis

After the data collection had been completed, the scale development assumptions of the Rasch model were tested, with unidimensionality and local independence being the two basic assumptions. Wright (1996) stated that factor analysis should test unidimensionality as an assumption in the Rasch model. In this first phase of the current study, we aimed to develop a single form for high school and undergraduate students. Based on this aim, Explanatory Factor Analysis (EFA) was performed on the data of 658 participants, without separating them according to educational level (i.e., high school and undergraduate students) to test unidimensionality. However, the factor structure of the high school and undergraduate student level indicated significant differences in terms of the number of items, factor loads, total-explained variance, and afterwards in producing distorted results in model-data fit. At this stage, we decided that test-taking strategies indicate dissimilar constructs at the high school and undergraduate level. Therefore, the subsequent analyses were conducted as two separate participant groups to test the validity of two separate scales, i.e., a high school form and an undergraduate form.

The scree-plot graphs (see Appendix 1 and Appendix 3) were used to determine the scale factors. Both forms of the scale were shown to have a one-dimensional structure, and factor loadings (see Appendix 2 and Appendix 4) were considered in deciding on the items included in both forms. According to Tabachnick and Fidell (2007) and Kline (2011), factor loads should be at least .32 to be included. In the current study, the .32 value was used to determine when items were included in the scale. EFA proofs and Martin-Löf test results were used for unidimensionality. Tennant and Conaghan (2007) suggested using inter-item residual correlation values to meet local independence, which is an assumption of the Rasch model. In the current study, we used a .40 value in analysing residual correlations between items.

Reliability was evaluated using the Person Separation Index (PSI) from the Rasch analysis. This is similar to coefficient alpha, but uses the metric latent trait in place of the summed score. The literature suggests that a PSI value of .7 or above reflects consistency (Tennant & Conaghan, 2007). After the Rasch model assumptions had been tested, estimates were made regarding PCM. The calculation used to assess the probability of getting an *x*-score from Item *j* of Student *i* is given in Equation 1.

$$\boldsymbol{P_{ijx} = \frac{exp \sum_{k=0}^{x}(\theta_i - \beta_{jk})}{\sum_{k=0}^{m} exp \sum_{t=0}^{k}(\theta_i - \beta_{jt})}} \quad \textbf{(1)}$$

PCM has an individual parameter θ and an item parameter β. The β parameter is defined as the 'step difficulty', which describes a student's successful completion and then moving on to the next step. The 'step difficulty' parameter is also known as the 'category intersection' parameter. Consequently, the step difficulty parameter was defined as the difficulty of choosing one response category over another response category. In PCM, the step difficulty parameters are one less than the item category number. For example, there would be three-step difficulty parameters for an item with four categories.
Insignificance of chi-square fit statistics is an indicator of item-model fit in PCM. Chi-square statistics are based on the difference between expected and observed values at different trait levels. In the current study, considering the Bonferroni correction, the .002 level was used to fit the item model (Bland & Altman, 1995). RStudio and R4.0.3 software with norm, mice, mnormt, psych, classInt, and eRm packages were used for the PCM estimates. IBM SPSS Statistics version 20 was used to process the data for the EFA and other analyses.

The 20-item Test-Taking Strategies Scale developed by Bıçak (2013) was used as the criterion reference to measure the validity of the scales. The developed forms and the Test-Taking Strategies Scale (Bıçak, 2013) were applied to 49 high school students for the high school form and 61 university students for the undergraduate form. Since the data was not normally distributed, the Spearman-Brown rank-order correlation coefficient was calculated for the correlation index.

Results and Discussion

Validity and Reliability Measures of High School Form

EFA was conducted to test unidimensionality, one of PCM's assumptions. The KMO value was found to be .95, and the Bartlett sphericity test result was significant (χ^2 = 8089,89; SD = 990; p = .000). Considering these results, we determined that the data was a good fit for factor analysis. It was revealed that the scale consisted of 47 items

within a one-dimensional structure. The factor loading values of two items were excluded from the scale since they were lower than .32. When the scree-graph in Appendix 1 is examined, it can be seen that the 45-item scale has a single dominant factor. Factor loadings of the items and their contributions to common variance are also presented in Appendix 1.

In testing the local independence, residual correlations between items were examined. The residual correlation value between the ninth and 10th items was determined to be .43. In examining these items, it was found that the ninth item ('I plan how I will use the time in relation to the whole test lesson') and the tenth item ('I try to estimate how much time I have available for each item') measured similar features and thus interfered with local independence. For this reason, it was decided to retain the ninth item, which is both more comprehensible and has a higher factor loading, while the 10th item was retained.
After testing the assumptions, the analysis of the remaining items in the scale was carried out according to PCM. It was determined that 17 of the 44 items did not show item-model compatibility. According to the Martin-Löf test statistic result for 27 items, no significant difference existed between the expected and observed values (LR-value: 796.518, p =.99). This result formed the second proof of unidimensionality. Item-model fit values for the remaining 27 items are presented in Table 2.

Table 2. PCM Item-Model Fit Indexes-High School Form

Item No	X^2	p	Outfit MS	Infit MS	Item No	X^2	p	Outfit MS	Infit MS
M8	328.501	.345*	1.027	0.774	M27	367.895	.031*	1.150	1.166
M9	366.095	.035*	1.144	1.048	M28	273.724	.968*	0.855	0.864
M11	389.714	.004*	1.218	1.182	M29	306.347	.685*	0.957	0.968
M12	324.365	.406*	1.014	1.044	M31	356.310	.074*	1.113	1.058
M14	391.029	.004*	1.222	1.147	M36	346.091	.142*	1.082	0.910
M16	341.149	.188*	1.066	1.091	M37	359.497	.059*	1.123	1.043
M17	246.575	.999*	0.771	0.768	M39	326.755	.370*	1.021	0.995
M18	286.628	.903*	0.896	0.939	M40	266.714	.985*	0.833	0.852
M19	326.348	.376*	1.020	1.043	M41	236.168	1.000*	0.738	0.786
M21	319.902	.475*	1.000	1.057	M42	267.441	.984*	0.836	0.888
M22	312.095	.598*	0.975	0.970	M43	311.057	.614*	0.972	0.942
M24	348.358	.124*	1.089	0.842	M45	228.121	1.000*	0.713	0.772
M25	367.548	.032*	1.149	1.129	M46	266.279	.986*	0.832	0.866
M26	343.724	.163*	1.074	1.000					

*p > .002

As summarised in Table 2, all 27 items showed item-model fit. Convenient quantitative measures of fit discrepancy are mean-square residual summary statistics, such as Outfit and Infit. These statistics have an expectation of 1.0, and range from 0 to infinity. Mean-squares greater than 1.0 indicate underfit to the Rasch model, i.e., data less predictable than the model expects. Mean-squares less than 1.0 indicate overfit to the Rasch model, i.e., data more predictable than the model expects. However, the reasonable ranges for Outfit and Infit for rating scales is considered to be 0.6-1.4 (Wright, 1996). According to Table 2, all values were within the 0.6-1.4 range. Item parameters calculated within PCM for 27 items are presented in Table 3.

Table 3. Item Parameters of High School Form

Item No	Location	b1	b2	b3	b4	Item No	Location	b1	b2	b3	b4
M8	0.100	-0.725	-0.348	0.369	1.106	M27	0.728	-0.110	0.365	1.407	1.250
M9	0.687	-0.262	0.421	0.943	1.645	M28	0.222	-0.711	0.209	0.620	0.769
M11	0.335	-0.509	0.175	0.840	0.835	M29	0.832	-0.681	0.883	1.290	1.835
M12	0.487	-0.267	0.177	0.542	1.496	M31	0.568	-0.458	0.473	0.748	1.508
M14	0.688	0.146	0.302	1.205	1.100	M36	-0.097	-0.985	-0.607	0.559	0.643
M16	0.180	-1.477	-0.299	0.908	1.586	M37	0.415	-0.432	0.326	0.793	0.974
M17	-0.195	-1.568	-0.518	0.169	1.135	M39	0.254	-1.296	0.321	0.522	1.469
M18	0.296	-0.398	-0.016	0.603	0.994	M40	-0.082	-1.511	-0.172	0.408	0.949
M19	0.167	-0.999	-0.207	0.610	1.266	M41	0.116	-0.736	-0.183	0.386	0.997
M21	0.582	-0.290	0.272	0.944	1.399	M42	0.157	-0.497	-0.008	0.382	0.751

Item No	Location	b1	b2	b3	b4	Item No	Location	b1	b2	b3	b4
M22	0.443	-0.366	0.208	0.462	1.468	M43	0.504	-0.447	0.019	0.636	1.806
M24	-0.155	-0.512	-0.260	0.078	0.074	M45	0.161	-0.636	-0.172	0.632	0.820
M25	0.352	-0.714	0.100	0.976	1.044	M46	0.478	-0.534	-0.017	0.857	1.605
M26	0.089	-1.061	-0.125	0.337	1.205						

As Table 3 shows, there were no disordered thresholds. As all of the items were polytomous, an analysis was conducted of each category's ordering. The issue here is whether the transition from a lower to a higher response category within an item was consistent with increases in the underlying trait. The scale's reliability was examined using Person Separation Index (PSI), which is equivalent to Cronbach's alpha, but has a linear transformation regarding the Rasch model. Tennant and Conaghan (2007) suggested that a coefficient score above .70 proves the consistency of a scale, and the PSI coefficient for the current study was calculated as .93. The correlation between high school form scores and criterion scale scores (Bıçak, 2013) was calculated to be 0.689 ($p < .01$). This mean correlation is evidence that the scales measure similar constructs. The result can also be interpreted as the degree of criterion-related validity.

Validity and Reliability Measures of Undergraduate Form

EFA was conducted to test the unidimensionality of the scale, which is one of PCM's assumptions. The KMO value was found to be .87 and the Bartlett sphericity test result was shown to be significant ($\chi^2 = 3771{,}149$; $SD = 561$; $p = .000$). Considering these findings, we determined that the data was well-fitted for factor analysis. It was revealed that the scale consisted of 47 items within a one-dimensional structure. The factor loading values of 13 items were excluded from the scale because they were lower than .32. When the scree-graph in Appendix 2 is examined, it can be seen that the 34-item scale has a single dominant factor. Factor loadings of the items and their contributions to common variance are also presented in Appendix 2.

In testing the local independence, residual correlations between the items were examined. The residual correlation value between the ninth and 10th items was determined to be .41. Examination of these items revealed that the ninth item ('I plan how I will use the time in relation to the whole test lesson') and the tenth item ('I try to estimate how much time I have available for each item') measure similar features and therefore interfere with local independence.For this reason, it was decided to retain the ninth item, which is both more understandable and has a higher factor load, whilst the 10th item was excluded from the scale.

After testing the assumptions, the analysis of the remaining items in the scale was carried out according to PCM. It was determined that 15 of the 33 items did not show item-model compatibility. According to the Martin-Löf test statistic result for 18 items, it was revealed that there was no significant difference established between the expected and observed values (LR-value: 399.42, $p = 1.000$). This result is considered as a second proof of unidimensionality. The item-model fit values for the remaining 18 items are presented in Table 4.

Table 4. PCM item-model fit indexes of undergraduate form

Item No	X^2	p	Outfit MS	Infit MS	Item No	X^2	p	Outfit MS	Infit MS
M5	336.609	.419	1.024	0.881	M38	365.218	.101	1.156	1.008
M8	318.322	.696	0.941	0.955	M40	404.040	.004	1.216	1.006
M9	368.928	.079	0.989	0.971	M41	326.974	.568	0.982	0.898
M17	300.465	.892	1.050	1.021	M42	340.701	.359	1.025	0.955
M18	368.169	.084	0.897	0.889	M43	286.451	.966	0.879	0.845
M19	332.427	.483	1.144	1.089	M44	265.771	.997	0.789	0.754
M22	371.701	.066	0.998	0.989	M45	260.580	.999	0.794	0.751
M28	298.103	.909	1.131	1.047	M46	328.430	.545	1.019	0.935
M29	387.678	.019	0.885	0.903	M47	406.961	.003	1.189	1.089

*$p > .002$

As summarised in Table 4, all 18 items showed item-model fit. Convenient quantitative measures of fit discrepancy are mean-square residual summary statistics, such as Outfit and Infit. These statistics have an expectation of 1.0, and range from 0 to infinity. Mean-squares greater than 1.0 indicate underfit to the Rasch model, i.e., data less predictable than the model expects. Mean-squares less than 1.0 indicate overfit to the Rasch model, i.e., data more predictable than the model expects. However, reasonable ranges for Outfit and Infit for

rating scales are suggested to be 0.6-1.4. (Wright, 1996). According to Table 4, all of the values are within the 0.6-1.4 range. The item parameters calculated within PCM for the 18 items are presented in Table 5.

Table 5. Item parameters of undergraduate form

Item No	Location	b1	b2	b3	b4	Item No	Location	b1	b2	b3	b4
M5	-0.662	-2.022	-0.534	-0.412	0.319	M38	-0.309	-1.627	-0.210	-0.020	0.619
M8	-0.181	-1.664	-0.290	-0.174	1.403	M40	-0.086	-1.621	-0.504	0.034	1.745
M9	0.561	-0.744	0.337	0.830	1.820	M41	0.586	-0.282	-0.001	0.941	1.684
M17	0.403	-0.515	-0.058	1.783	1.830	M42	0.563	-0.063	0.204	0.279	1.831
M18	0.779	-0.415	0.741	1.185	1.606	M43	0.616	-0.652	-0.575	0.307	2.078
M19	0.448	-0.539	-0.329	0.738	1.923	M44	0.567	-0.403	-0.355	0.250	1.969
M22	0.727	-0.758	0.850	0.857	1.960	M45	0.329	-0.372	-0.319	0.446	1.562
M28	0.049	-1.267	-0.419	0.231	1.649	M46	0.618	-0.600	0.115	0.769	2.187
M29	1.069	0.353	0.502	1.161	2.262	M47	0.577	-0.212	-0.038	1.302	1.502

As Table 5 shows, there were no disordered thresholds. As all of the items were polytomous, an analysis was undertaken of the ordering of each category. The issue here was whether or not the transition from a lower to a higher response category within an item was consistent with an increase in the underlying trait. The scale's reliability was examined using the PSI, which is equivalent to Cronbach's alpha, but has a linear transformation from the Rasch model. In the current study, the PSI value was calculated as .88. The correlation between undergraduate form scores and criterion scale (Bıçak, 2013) scores were found as 0.805 ($p>.01$). This high correlation emphasize the similarity of the constructs measured by the scale developed for similar purposes. The finding constitutes important evidence for criterion referenced validity.

Discussion and Conclusion

Examination and test scores play an important role in modern life, and test-taking strategies are considered an important factor affecting test scores. However, attempts to measure test-taking strategies are seen as relatively new, and there is no complete agreement, as yet, from a theoretical perspective (Bıçak, 2013; Cohen, 2007; Dodeen, 2008; Hong et al., 2006; Pehlivan & Kutlu, 2014; Peng et al., 2014). The effect of test-taking strategies on different psychological characteristics related to testing has been investigated in the literature (Beidel et al., 1999; Bruch, 1981; Chittooran & Miles, 2001; Dodeen, 2014; Dodeen et al., 2014; Kesselman-Turkel & Peterson, 2004; Peng et al., 2014). On the other hand, test results should be calculated without using an test-taking strategy (Smith, 2017) as a confounding psychological feature unrelated to the measured structure. As a result of the current research, valid, reliable, and up-to-date scales measuring test-taking strategies were developed for different grade levels. The developed scales are expected to contribute to the field and to their application as they have been shown to make assessments with a high degree of validity.

The study was initiated to develop a scale focusing on the 17-22 year old student age group, which frequently encounter exams during their education. However, in the validity analysis of the research data, it was observed that test-taking strategies at the high school and undergraduate level showed significant differences in the psychological construct. Aside from the purpose of the current study, an additional finding was that test strategy structures differed according to the schooling level. Therefore, in the current study we developed both a high school form (consisting of 27 items) and a university form (consisting of 18 items) so as to measure students' test-taking strategies. The developed scales are 5-point, Likert type instruments, with no reverse scoring item in either scale. The minimum score for the high school form is 27 and the maximum score is 135. The minimum score of the undergraduate form is 18, and the maximum score is 90. As the scores approach the maximum score, the students' level of using test-taking strategies increases.

Literature focusing on determining test strategies include qualitative research which describe individual's response processes (Hong et al, 2006; Chittooran & Miles, 2001), questionnaires (Hong & Peng, 2004; Peng et al, 2014) and contemporary research examine scale development (Bıçak, 2013; Dodeen, 2008). With this study, two scales have been developed to determine test strategies; that have not been well-defined construct in the literature. Two developed scales differ from similar by IRT based validity studies and the test items are structured according to the temporal dimension of the test as "pre-test, during test and post-test" indicators. In particular, this research will shed light on future research by presenting two separate forms for different educational levels.

Exams that students encounter in high school and university may differ in terms of practice, the skills they test, and the associated stakes (Boud & Falchikov, 2007). In the case of Turkey, the grades obtained throughout high

school education and the results of university entrance exams are used in decision-making to enter higher education; in other words, to commence education for a profession (Abrams, 2004; Flitcroft et al., 2017). The test-taking strategies that high school students may employ are varied and numerous to be successful in such high-stake exams that will ultimately shape their lives from that point onwards. In supporting these high-risk exam behaviours, several scale items were included in the high school form regarding the duration of the exam (Item 36), caring about response control (Item 37), and efforts to prove what they know (Item 27 and Item 31). In high-stake tests, multiple-choice items are predominantly included. It is noteworthy that some items (e.g., Item 24, Item 25, and Item 26) that refer to multiple-choice items in the high school form are not included in the university form. In higher education, educational goals focus more on high-level skills and specialisation (Fallows et al., 2000), and measurement is conducted accordingly. The number of test strategies that can be used in examinations for high-level skills such as making a product, a performance, an evaluation, and a synthesis, and their effects on the measurement result can be somewhat limited. Therefore, the university form consisted of nine fewer items than the high school form.

The Rasch model in psychological tests that use scoring with grading totals, such as Likert-type scales, is still considered to be quite new. It is known that more valid and reliable results are provided due to the advantages of IRT and the Rasch model. The PCM helps by comparing different versions of scales to decide which form provides the most valid and reliable results for the construct being measured. Therefore, it is possible to use the results of two forms to monitor and improve measures until they reach the level of measurement accuracy required for decision making (Van Zile-Tamsen, 2017). The validity and reliability of the two forms developed in the current study were supported by precise estimates.

References

Abrams, L. M. (2004). *Teachers' views on high-stakes testing: Implications for the classroom policy brief*. Arizona State University, Education Policy Studies Laboratory. http://edpolicylab.orghttp//edpolicylab.org

American Psychological Association, American Educational Research Education, & National Council on Measurement in Standards. (1966). *Standards for educational and psychological tests and manuals*. APA.

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal*, *75*(4), 460-472. https://doi.org/10.2307/329495

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university.

Baykul, Y. (2010). *Eğitim ve psikolojide ölçme ve değerlendirme [Measurement and evaluation in education and psychology]*. Pegem Akademi.

Beidel, D. C., Turner, S. M., & Taylor-Ferreira, J. C. (1999). Teaching study skills and test-taking strategies to elementary school students. *Behavior Modification*, *23*(4), 630-646. https://doi.org/10.1177%2F0145445599234007

Bıçak, B. (2013). Teste hazırlık ve test yanıtlama stratejileri ölçeği [Scale for test preparation and test taking strategies]. *Kuram ve Uygulamada Eğitim Bilimleri*, *13*(1), 273-289.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *Bmj*, *10*(6973), 170.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Boud, D., & Falchikov, N. (2007). *Rethinking assessment in higher education: Learning for the longer term*. Routledge.

Boyd, R. T. C. (1989). Improving your test-taking skills. *Practical Assessment, Research, and Evaluation*, *1*(1), Article 2. https://doi.org/https://doi.org/10.7275/ypnv-6b10

Bruch, M. A. (1981). Relationship of test-taking strategies to test anxiety and performance: Toward a task analysis of examination behavior. *Cognitive Therapy and Research*, *5*(1), 41–56.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. British Council.

Chittooran, M. M., & Miles, D. D. (2001). *Test-taking skills for multiple-choice formats: Implications for school psychologists.* Saint Louis University.

Cohen, A. D. (2007). The coming of age for research on test-taking strategies. In *Language Testing Reconsidered* (pp. 89–112). University of Ottawa Press.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Cengage Learning.

Demirci, İ., & Erden, S. (2016). Bilişsel davranışçı yaklaşıma dayalı grupla psikolojik danışma uygulamasının 8. sınıf öğrencilerinin sınav kaygısına etkisi [The effect of cognitive-behavioral group counseling on test anxiety of eight-grade elementary school students]. *Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi*, *43*(43), 67. https://doi.org/10.15285/ebd.51646

Demirtaşli, N., Yalçin, S., & Ayan, C. (2016). Ölçme ve değerlendirme dersine yönelik tutum ölçeğinin madde tepki kuramına dayalı olarak geliştirilmesi [The Development of IRT Based Attitude Scale towards Educational Measurement Course]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *7*(1), 133–144. https://doi.org/10.21031/epod.43804

Dodeen, H. (2008). Assessing test-taking strategies of university students: Developing a scale and estimating its psychometric indices. *Assessment and Evaluation in Higher Education*, *33*(4), 409–419. https://doi.org/10.1080/02602930701562874

Dodeen, H. (2014). Test-related characteristics of 0UAEU students: Test-anxiety, test- taking skills, guessing, attitudes toward tests, and cheating. *Journal of Faculty of Education*, *26*(J2009), 31–66.

Dodeen, H. (2015). Teaching test-taking strategies: importance and techniques. *Psychology Research*, *5*(2), 108–113. https://doi.org/10.17265/2159-5542/2015.02.003

Dodeen, H. M., Abdelfattah, F., & Alshumrani, S. (2014). Test-taking skills of secondary students: The relationship with motivation, attitudes, anxiety and attitudes towards tests. *South African Journal of Education*, *34*(2), Article 866. https://doi.org/10.15700/201412071153

Educational Testing Service. (2001). *Graduate Record Examinations psychology practice book*. Educational Testing Service.

Embretson, S., & Reise, S. P. (2000). *Item Response Theory for Psychologists* (1st ed.). Erlbaum.

Engelhard, G., & Wind, S. A. (2017). *Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments*. Routledge. https://doi.org/10.4324/9781315766829

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme [Measurement and scale development in psychology]*. Pegem Akademi.

Fallows, S. J., Fallows, S., & Steven, C. (2000). *Integrating key skills in higher education: Employability, transferable skills, and learning for life*. Psychology Press.

Flitcroft, D., Woods, K., & Putwain, D. W. (2017). Developing school practice in preparing students for high-stake examinations in English and Mathematics. *Educational and Child Psychology*, *34*(3), 7–19.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet rasch model. *Eurasian Journal of Educational Research*, *55*, 73–90. https://doi.org/10.14689/ejer.2014.55.5

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hong, E., Sas, M., & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *Journal of Educational Research*, *99*(3), 144–155. https://doi.org/10.3200/JOER.99.3.144-155

İlhan, M., & Güler, N. (2018). Likert tipi ölçeklerde Rasch modelinin kullanımı: olumsuz değerlendirilme korkusu ölçeği-öğrenci formu (ODKÖ-ÖF) üzerinde bir uygulama [The use of Rasch model in Likert types scales: An application on the fear of negative evaluation scale-student form. *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, *8*(4), 756-775.

Kartal, K., & Dirlik, M. (2016). Historical development of the concept of validity and the most preferred technique of reliability: Cronbach alpha coefficient. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *16*(4), 1865-1879.

Kaya Uyanık, G., Güler, N., Taşdelen Teker, G., & Demir, S. (2019). Fen bilimleri dersi etkinliklerinin Çok Yüzeyli Rasch Modeliyle analizi [The analysıs of elementary science education course actıvıtıes through Many-Facet Rasch Model]. *Kastamonu Eğitim Dergisi*, *27*(1), 139–150. https://doi.org/10.24106/kefdergi.2417

Kesselman-Turkel, J., & Peterson, F. (2004). *Test-taking strategies*. University of Wisconsin.

Kizilkaya, G., & Aşkar, P. (2009). Problem çözmeye yönelik yansitici düşünme becerisi ölçeğinin geliştirilmesi [The development of a reflective thinking skill scale towards problem solving]. *Egitim ve Bilim*, *34*(154), 82-92.

Koch, W. R. and Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, 2(4), 335-357.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

Kutlu, Ö., Yıldırım, Ö., & Bilican, S. (2009). Öğretmenlerin dereceli puanlama anahtarlarına ilişkin tutum ölçeği geliştirme çalışması [Study of attıtudes scale development aımed at scorıng rubrıcs for prımary school teachers]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *6*(2), 76-88.

Lewandowski, L., Gathje, R. A., Lovett, B. J., & Gordon, M. (2013). Test-taking skills in college students with and without ADHD. *Journal of Psychoeducational Assessment*, *31*(1), 41–52. https://doi.org/10.1177/0734282912446304

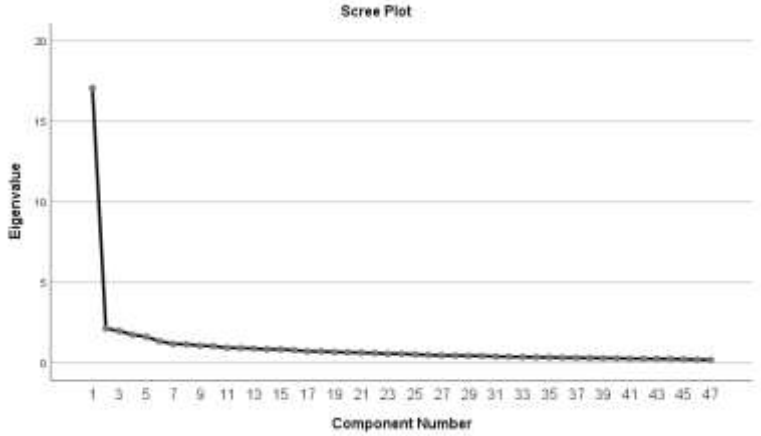
Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Mesa.

Pehlivan, E. B., & Kutlu, Ö. (2014). Türkçe test maddelerinde yanıtlama davranışlarının incelenmesi [Investigation of answering behaviour in Turkish test]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *5*(1), 61–71. https://doi.org/10.21031/epod.20130

Peng, Y., Hong, E., & Mason, E. (2014). Motivational and cognitive test-taking strategies and their influence on test performance in mathematics. *Educational Research and Evaluation*, *20*(5), 366–385.

https://doi.org/10.1080/13803611.2014.966115
Primi, R., Silvia, P. J., Jauk, E., & Mathias, B. (2019). Applying Many-Facet Rasch Modeling in the Assessment of Creativity. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 176–186. https://doi.org/10.1037/aca0000230
Roderer, T., & Roebers, C. M. (2014). Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance. *Metacognition Learning*, *9*, 1–23. https://doi.org/10.1007/s11409-013-9109-4
Rozakis, L. (2003). *Test-taking strategies and study skills for the utterly confused*. McGraw-Hill.
Şenel, S., & Kutlu, Ö. (2018). Comparison of two test methods for VIS: paper-pencil test and CAT. *European Journal of Special Needs Education*, *33*(5), 631–645. https://doi.org/10.1080/08856257.2017.1391014
Şenel, S., & Şenel, H. C. (2018). Bilgisayar tabanlı testlerde evrensel tasarım: Özel gereksinimli öğrenciler için düzenlemeler [Universal design for computer-based testing: Accommodations for students with special needs]. In S. Dinçer (Ed.), *Değişen dünyada eğitim [Education in versatile world]* (1st ed., pp. 113-124). Pegem Akademi. https://doi.org/10.14527/9786052412480.08
Smith, M. D. (2017). Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes? *American Educational Research Journal*, *54*(6), 1256-1287. https://doi.org/10.3102/0002831217717949
Tabachnick B G & Fidell L S (2007). *Using multivariate statistics*. Boston, Pearson Education, Inc.
Tekin, H. (1996). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (9th ed.). Yargı.
Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358-1362.
Therrien, W. J., Hughes, C., Kapelski, C., & Mokhtari, K. (2009). Effectiveness of a test-taking strategy on achievement in essay tests for students with learning disabilities. *Journal of Learning Disabilities*, *42*(1), 14–23. https://doi.org/10.1177/0022219408326218
Ün Açıkgöz, K. (2003). *Etkili öğrenme ve öğretme [Effective learning and teaching]* (6th ed.). Biliş.
Van Zile-Tamsen, C. (2017). Using Rasch Analysis to Inform Rating Scale Development. *Research in Higher Education*, *58*(8), 922-933. https://doi.org/10.1007/s11162-017-9448-0
Wongpakaran, N., Wongpakaran, T., Pinyopornpanish, M., Simcharoen, S., Suradom, C., Varnado, P., & Kuntawong, P. (2020). Development and validation of a 6-item Revised UCLA Loneliness Scale (RULS-6) using Rasch analysis. *British Journal of Health Psychology*, *25*(2), 233-256. https://doi.org/10.1111/bjhp.12404
Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, *10*(3), 509-511.
Yaşar, M., & Aybek, E. C. (2019). A resilience scale development for university students: Validity and reliability study based on item response theory. *Elementary Education Online*, *18*(4), 1687-1699. https://doi.org/10.17051/ilkonline.2019.635031
Yıldırım, A., Doğanay, A., & Türkoğlu, A. (2000). *Okulda başarı için ders çalışma ve öğrenme yöntemleri [Study and learning methods for success in school]*. Seçkin Academy.

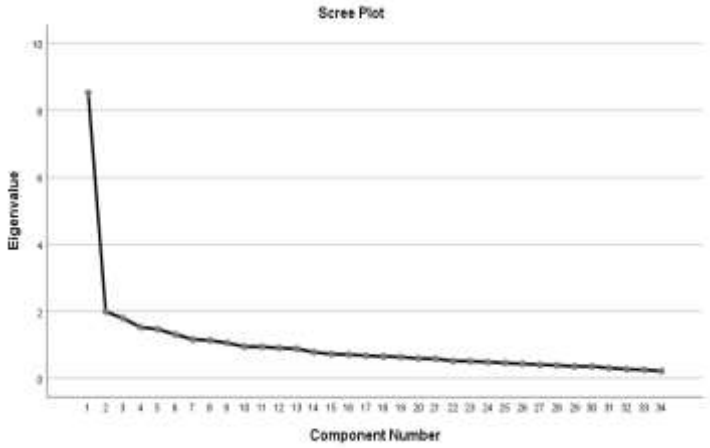
Appendix 1. Scree-Plot Graph of Secondary Education Data



Appendix 2. EFA Results of High School Data: Factor Loadings and Communalities

Item number	Factor loading	Communality	Item number	Factor loading	Communality
M2	.37	.13	M25	.56	.31
M3	.34	.11	M26	.61	.37
M4	.56	.32	M27	.53	.28
M5	.69	.48	M28	.68	.47
M6	.57	.33	M29	.61	.37
M7	.53	.28	M30	.53	.28
M8	.75	.56	M31	.62	.38
M9	.61	.38	M32	.51	.26
M10	.57	.33	M33	.65	.43
M11	.56	.31	M34	.44	.19
M12	.62	.38	M36	.69	.47
M13	.56	.31	M37	.62	.38
M14	.56	.31	M38	.79	.62
M15	.54	.29	M39	.63	.39
M16	.55	.31	M40	.70	.49
M17	.73	.53	M41	.75	.56
M18	.68	.46	M42	.71	.50
M19	.60	.36	M43	.67	.45
M20	.57	.32	M44	.68	.46
M21	.58	.34	M45	.76	.58
M22	.64	.41	M46	.69	.48
M23	.43	.19	M47	.52	.27
M24	.70	.49			

Appendix 3. Scree-Plot Graph of Undergraduate Data



Appendix 4. EFA Results of Undergraduate Data: Factor Loadings and Communalities

Item number	Factor loading	Communality	Item number	Factor loading	Communality
M3	.40	.16	M28	.58	.34
M4	.40	.16	M29	.52	.27
M5	.52	.27	M30	.39	.16
M8	.54	.29	M31	.45	.20
M9	.56	.32	M33	.43	.19
M10	.54	.29	M36	.40	.16
M12	.41	.17	M37	.43	.18
M13	.32	.10	M38	.50	.25
M16	.44	.19	M39	.47	.22
M17	.59	.35	M40	.52	.27
M18	.56	.31	M41	.59	.35
M19	.55	.30	M42	.54	.29
M20	.36	.13	M43	.61	.37
M21	.39	.16	M44	.68	.46
M22	.54	.30	M45	.69	.47
M24	.35	.12	M46	.59	.35
M26	.34	.12	M47	.52	.27

Appendix 5. Test Strategies Pilot Scale Items and Included in Scales (in Turkish)

Item No	*OE* abbreviation was used for Open-ended test items. *MC* abbreviation was used for Multiple-Choice test items.	High School Form	University Form
M5	Sınava gerekli tüm materyalleri getiririm.	-	Included
M8	Sınav açıklamalarını dikkatli biçimde okurum.	Included	Included
M9	Toplam sınav süresine göre, süreyi nasıl kullanacağımı planlarım.	Included	Included
M11	Yanıtlamaya başlamadan önce sınav kâğıdındaki tüm sorulara hızlıca göz atarım.	Included	-
M12	Sınav kâğıdının boş yerlerine soruları yanıtlarken yararlanabileceğim notları (formül, anahtar kelime vb.) yazarım.	Included	-
M14	Sınava en kolay olduğunu düşündüğüm sorudan başlarım.	Included	-
M16	Soruları yanıtlarken sorunun kökünü birden çok kez okurum.	Included	-
M17	Soruları yanıtlarken sorunun köküne (ne istendiğine) odaklanırım.	Included	Included
M18	Sorudaki anahtar sözcüklerin altını çizerim.	Included	Included
M19	Karmaşık soruları, kendi cümlelerimle zihnimde tekrar düzenlerim.	Included	Included
M21	Bir soruyu planladığım sürede yanıtlayamamışsam diğer soruya geçerim.	Included	-
M22	Her yanıtımdan sonra yanıtımı hızlıca kontrol ederim.	Included	Included
M24	(MC) Öncelikle kesinlikle yanlış olduğunu düşündüğüm seçenekleri elerim.	Included	-
M25	(MC) Soruları yanıtlandırırken diğerlerinden farklı görünen seçeneği elerim.	Included	-
M26	(MC) İki-üç seçenek arasında kaldığımda doğru yanıtı tahmin etmeye çalışırım.	Included	-
M27	(OE) Sorunun yanıtını bilmiyorsam, konu ile ilgili bildiğim her şeyi yazarım.	Included	-
M28	(OE) Yanıtımı yazmadan önce, yazacaklarımı zihnimde düzenlerim.	Included	Included
M29	(OE) Soruların altında, düzeltme / ekleme için bir miktar boşluk bırakırım.	Included	Included
M31	(OE) Bazı sorular için zamanım kalmazsa, yanıtların ana hatlarını yazarım.	Included	-
M36	Sınav süresini sonuna kadar kullanırım.	Included	-
M37	Tüm soruları yanıtlayamamış olsam bile son birkaç dakikamı, yanıtlarımı kontrol etmeye ayırırım.	Included	-
M38	Zamanım kalırsa, yanıtlarımı kontrol ederim.	-	Included
M39	Sınav anında, sınav sonucundan çok sınava odaklanırım.	Included	-
M40	Yanıtı bilmiyorsam, akıllıca tahminlerde bulunmaya çalışırım.	Included	Included
M41	Sınav sonrasında doğrularımı, yanlışlarımı, eksiklerimi ve hatalarımı kontrol ederim.	Included	Included
M42	Sınav sonrasında diğer öğrencilerin veya ders sorumlusunun yaptığı değerlendirmeleri dikkatle dinlerim.	Included	Included
M43	Sınav anındaki çabamı objektif olarak değerlendiririm.	Included	Included
M44	Puanımı düşüren nedenleri düşünürüm.	-	Included
M45	Bir sonraki sınavda performansımı nasıl artırabileceğimi düşünürüm.	Included	Included
M46	Sınav sonucuna göre, gerekirse sınava hazırlık yöntemlerimde değişiklik yaparım.	Included	Included
M47	Sınavım iyi geçerse kendimi ödüllendiririm.	-	Included

(-) excluded