Education and Science

Vol 39 (2014) No 174 201-213

Psychometric Properties of Diagnostic Branched Tree

Ümit Çelen [1]

Abstract

Various changes are observed in assessment methods and tools along with the adoption of constructivist teaching and learning approach. One of the assessment tools proposed by the new curriculum is the diagnostic branched tree. Data for this study which was conducted to identify the psychometric properties of the tool were obtained from three working groups. Common denominators of the total of 525 students participating in the study were that they were all teacher candidates and they took assessment and evaluation course in 2012-2013 academic year. Data were collected via diagnostic branched tree developed by the researcher to evaluate achievement in the measurement and evaluation class, true-false test and multiple choice achievement tests. Results show that diagnostic branched tree also includes the disadvantages of the true-false tests, does not have high reliability and validity and it is not as "diagnostic" as the name implies.

Keywords

Diagnostic branched tree
True-false tests
Validity
Reliability

Article Info

Received: 04.02.2013
Accepted: 04.29.2014
Online Published: 08.06.2014

DOI: 10.15390/EB.2014.2630

Introduction

Measurement and evaluation are used in education for various purposes such as student selection for specific school programs, recognition of student characteristics and identification of readiness, learning/competence levels and achievement of lesson objectives. The quality of teaching and training is affected by the quality of the assessment process and a firm connection is believed to exist between teaching process and evaluation. Based on the fact that teaching and evaluation are interrelated, teachers do not only focus on grading students but use measurement and evaluation to monitor academic development of students and to feedback to the process by paying attention to use various tools and methods in measurement and evaluation (Çıkrıkçı-Demirtaşlı, 2012).

Since 2005-2006 academic year, Ministry of National Education (MoNE) in Turkey adopted constructivism that was developed in 2004 as the fundamental teaching and training approach in curriculums and this approach was also reflected in measurement and evaluation processes. Tools that focus on process assessment and evaluation were proposed in the renewed curriculums in addition to classical measurement tools (MoNE, 2005).

It is possible to claim that constructive approach is based on ancient philosophical roots although it was raised in the first half of the 1960s by Bruner. According to Dubs (1993), behavioral and cognitive approaches place meaning and knowledge on an objective basis whereas constructivist approach regards knowledge to be structured with internal processes and individually. There are crucial differences between behavioral and cognitive approaches and constructivist approach

[1] Amasya University, Faculty of Education, Educational Sciences, Department of Measurement and Evaluation, Turkey, umitcelen@yahoo.com

regarding what knowledge and truth are and where they are found (Cited in Şimşek, 2004). Constructivism has recently been an important orientation and topic of discussion in educational sciences literature. "This approach, which denies objective knowledge either completely or to a great extent, is based on agreement, cooperation, culture and the variability, provisionality and the contingency of knowledge or proposes subjectivity and relativity as indispensable principles. In this sense, it is underestimated as much as it is adopted" (Şimşek, 2004: 115).

Student-centered teaching methods based on constructivist approaches are adopted in the new curriculums since 2005-2006 academic year. Lesson contents, teaching methods, tools and equipment used in lessons and measurement and evaluation methods have also changed along with the renewed curriculums (Gelbal and Kelecioğlu, 2007). One of the new assessment approaches that have become popular in educational environments is the performance based assessment. In performance based assessment, students are expected to undertake higher complexity tasks that require higher thinking skills instead of simple and plain tasks that call for lower level thinking skills. The purpose is to bring out competences in students such as creativity, problem solving, critical thinking and decision making and empathy and to identify the level of skill use and development (Kutlu, Doğan and Karakaya, 2008). Identification of student achievement in traditional approaches focuses more on product compared to teaching process and written and oral tests with multiple choice and short-answer formats are generally used. Assessment and evaluation in constructivist approach is a part of the process itself. Focusing on the process requires more frequent use of assessment and evaluation methods with different qualities (Gelbal and Kelecioğlu, 2007).

Linn and Gronlund (1995) stated that learning outputs that cannot be assessed through classical methods can be assessed with performance based assessment. In addition to paper and pencil tests, constructivist approach evaluates student performance from all aspects by observing student behavior and performance during the process, evaluating interest and attitude and including learners in the assessment process (Gelbal and Kelecioğlu, 2007).

According to Bachman (2002), performance assessment, sometimes called alternative or authentic assessment, has been very popular and a topic of discussion in many countries. It has opened up a new field for researchers to review the traditional outlook to characteristics that point to test quality such as validity. There are various assessment and evaluation methods and tools in the literature used in performance assessment processes. Student portfolios, performance tasks, projects, self-assessment forms, peer assessment forms, group assessment forms, observations, interviews, posters, presentations, rubrics, structured grids, diagnostic branched trees (DBT), word associations, concept maps and control lists are among these methods and tools (Özdemir, 2003). DBT, cited among tools that can be used in performance assessment, is included in the curriculums published by MoNE (2013). This tool is suggested for use in "information technologies" and "science and technology (4-5) curriculums and the use of the tool is explained by providing examples in "science and technology (6-8)" and "biology (9)" curriculums as well. The tools and methods to be used for assessment and evaluation in Biology (9) class are cited as tests that consist of questions in multiple choice, short-answer, matching, true-false and essay formats, performance assessment, portfolio assessment, research and projects, concept maps, structured grids, branched trees and self and peer assessment and assessment tools used in the framework of these techniques (rubrics and scoring scales etc.). The tool is explained in the Science and Technology (6-8) curriculums as follows (MoNE, 2013):

> *"DBT is one of the assessment tools that can be used to identify what students have learned or what they cannot have accomplished in a given subject. In this technique, students are asked to select the correct choice from among true and false statements in an order from basic statements to statements with more detail. Therefore, a branched tree made up of 8 or 16 selected statements is formed."*

"Branching Trees and Diagnostic Testing" by Johnstone et al. published in 1986 is the first article in which the term DBT was used in educational literature (Bahar, 2001). DBT is a tool that aims to identify knowledge patterns and misconceptions in cognitive structures via the true-false responses for associated proposition statements included in a tree graphic (Kocaarslan, 2012). Bahar et. al. (2009) argues that false associations, erroneous strategies and misconceptions in the minds of the responders are revealed via DBT. Kocaarslan (2012) states that DBT, developed to assess knowledge patterns in the cognitive states of responders and meaningful learning, is a suitable technique to allow using prior knowledge and experiences and forming associations among pieces of knowledge. The researcher also argues that chance effect ratio of 50% for selecting between true-false items decreases to 12.5%. There are findings stating that use of DBT increases student achievement (Şeyihoğlu and Erbaş, 2010) and success levels (Karahan, 2007).

The term "diagnostic" in the name of the tool means "to diagnose-teşhis etmek" in Turkish. The word "teşhis" comes from the Arabic root which means "understanding what or who, recognize, select" (TLA, 2013). The foundation of DBT which is supposed to diagnose is composed of true-false (TF) statements. TF statements are weak to diagnose areas which are miscomprehended by students (Tekin, 1996). This weakness is based on two reasons: First, the ratio of giving a correct answer to item by chance is 50%. A individual may correctly answer the item with a very high success rate such as 50% without the necessary quality that the item aims to assess. The second weakness is related to the probability that the responder may select the item as false for unrelated reasons and not knowing the true proposition involved. For instance, a student who selects false to the statement "The capital of France is Rome" can do it without knowing that the true answer is Paris indeed. In a 7-item and 15-item DBT, the responder needs to answer 3 and 4 items respectively. The probability of providing correct answers by chance effect is $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$ for 3 items and $\left(\frac{1}{2}\right)^4 = \frac{1}{16}$ for 4 items. The number of output branches is 8 or 16 in DBTs. Therefore, the chance for an individual to obtain the correct output without even reading any of the items is the same in 3 or 4-item TF tests.

Another weak aspect of TF items is their lack of assessment in higher level cognitive skills. The cognitive processes used in making decisions about the correctness-falseness of a proposition are at recognition-recall and comprehension level. A DBT presented by forming an associated set of items will be inadequate in assessing cognitive skills since it will be composed of TF items. Kutlu (2006) states that some classical test methods composed of multiple choice (MC), short-answer TF, matching and fill-in-the-blanks formats are insufficient to identify higher level cognitive levels such as problem solving, reading comprehension, critical thinking, analytical thinking, empathy, research, decision making, comprehending the importance of social history and creativity. Therefore, performance, portfolio and real life experiences based approaches are more beneficial to identify the level of student achievements foreseen in school programs. If the methods used in performance based approaches are defined as "methods used to assess learning outputs that cannot be assessed through classical methods" (Linn and Gronlund, 1995), then, the use of DBT among these tools is open to discussion.

Along with constructivist approach, studies on DBT mostly focus on teachers' knowledge levels about the tool and identification of the use of the tool in classes (Aydoğmuş and Coşkun Keskin, 2012; Karamustafaoğlu, Çağlak and Meşeci, 2012; Şaşmaz Ören, Ormancı and Evrekli, 2011; Çepni and Şenel Çoruhlu, 2010; Özdemir, 2010; Şenel Çoruhlu, Er Nas and Çepni, 2009; Okur, 2008; Pullu, 2008). The purpose of this study is to determine the psychometric properties of DBT by using an example that is developed for this aim.

Method

The study utilizes relational screening model (Karasar, 2003).The study aims to identify the psychometric properties of DBT and to compare the properties of DBT with those of TF and MC tests.

Working Group

Data were obtained from three different working groups attending Eskişehir Osmangazi University Faculty of Education and Faculty of Theology. The first working group (WG1) was composed of voluntary undergraduate students that attended the measurement and evaluation course during the fall semester of 2012-2013 academic year. Students in the second working group (WG2) attended the formation certificate program at the same university and took measurement and evaluation course in the same period. The third working group (WG3) was formed of undergraduate students attending measurement and evaluation course in the spring semester of 2012-2013 academic year. Table 1 presents the information regarding the distribution of the three working groups according to departments.

Table 1. Distribution of Students in Working Groups Based on Departments

Working Group 1			Working Group 2			Working Group 3		
Department	n	%	Graduated From	n	%	Department	n	%
PCG*	48	30.2	Turkish Language and Lit.	71	37.1	Primary Education Math. Teach.	66	37.9
Class. Teach.	91	57.2	Modern Turkish Dialects and Lit.	5	2.6	Primary Education Science Teach.	65	37.4
EREP**	20	12.6	Theology	19	9.9	CEIT***	29	16.7
Total	159	100.0	History	16	8.3	Other****	14	8.0
			Nursing	2	1.0	Total	174	100.0
			Mathematics	49	25.5			
			Physics	10	5.2			
			Chemistry	10	5.2			
			Biology	10	5.2			
			Total	192	100.0			

* Psychological Counseling and Guidance Program, ** Education of Religion and Ethics Program,
*** Computer Education And Instructional Technology, ****Students registered in the class from other departments.

Data Collection Tool

DBT: DBT used in the study was developed by the researcher. It was prepared in the framework of "Measurement and Evaluation" course which is a 3-credit course in undergraduate programs and 2-credit course in formation certificate programs. The tool consists of 15 TF items. The responder needs to answer 4 of these 15 items. The key answer to 8 of these items is "false" and key answer to 7 of these items is "true". The tool based on the key concepts related to assessment and evaluation includes items about measurement, evaluation, error and scales. Following the preparation of the tool, views of an expert on measurement and evaluation were sought regarding validity and adjustments were done based on these views.

Midterm: Midterm exam for the WG1 is a teacher prepared test composed of 25 items with 5 options for each item and was prepared by the researcher. The test included the subjects of importance of measurement, basic concepts in measurement and evaluation and errors in assessment and desired properties in measurement tools (reliability, validity and usefulness) to ensure content validity. Average test difficulty in the test in which no correction formula was applied was found to be 0.70; average point-biserial item discrimination 0.29, KR-20 internal consistency coefficient 0.54, arithmetic mean =17.56 and standard deviation was found to be S= 3.04.

Final Exam: The final exam WG2 took was a 40-item test with 5 options prepared by the researcher. In addition to the main exam topic mentioned above, the test included measurement tools used in education and statistical procedures implemented on the results of measurement. Average test difficulty in the test in which no correction formula was applied was found to be 0.70; average point-biserial item discrimination 0.29, KR-20 internal consistency coefficient 0.74, arithmetic mean =27.93 and standard deviation was found to be S= 5.04.

The study also utilized a 3-item interview form given to WG3. The three items in the form were related to their views on the efficiency of DBT to identify achievement and whether they thought of using the tool in the future when they started teaching.

Data Collection

Data from WG1 were obtained in the midterm exam and DBT implementation three weeks after the midterm. Students were asked to reach the output by first answering the test during the DBT implementation given in the class hour. Following the implementation, students were given the 15 items included in DBT in TF test format and they were asked to answer all the items.

Data from WG2 were obtained in the final exam. Two test forms (A and B) were generated by changing the placement of the items and one test form included DBT whereas the other form included 4 items that were on the correct output of the DBT.

When student scores are calculated, 30% of the midterm exam and 50% of the final exam are used. 20% of the scores related to development task is also added to their grade score.

Data from WG3 were obtained during the implementation undertaken one week after the topics in the framework of DBT were studied in the class. Students were given DBT form to answer and scoring was explained after the implementation. After calculating student scores based on the outputs found by students, the researcher explained all the items to students. Then the students were given interview forms regarding their views on the use of DBT.

Analysis of Data

All data obtained from the study was transferred to digital environment by using Microsoft Excel and analyses were done with the help of this program and SPSS 15.0. Point biserial correlation technique was used in the statistics calculated as discrimination power index. Since it was identified during the examination of correlations among test scores that the variables provided normal distribution assumption, Pearson product moment correlation technique was utilized. During item difficulty calculations for DBT, the difficulty of one item in the first section was calculated and later average scores for the section were used for the next three sections. Normality assumptions were checked and ensured in examining the differences between estimated scores obtained from DBT and obtained by changing the placement of two items in DBT and related t-test which was a parametric test was used. Stepwise method was used in the multiple linear regression analysis undertaken to observe whether DBT, TF test and estimated DBT scores predicted midterm exam scores. Questions to interview items were examined in terms of frequency and percentage distributions. Since normality assumptions were not ensured in examining the relationship between student sores and their thoughts on the sufficiency of DBT, Kruskal Wallis variance analysis was used.

Results

Table 2 presents the descriptive statistics obtained from conducting the test composed of DBT and 15 TF items in DBT to WG1. Students obtained an average score of 2.59 from DBT and 10.45 from TF test. Examination of average difficulties shows that TF test was easier for students however the difference is only 0.05. Correlation between two test scores was found to be 0.45. ınternal consistency for the four sections of DBT was calculated to be 0.03. This coefficient was found to be 0.37 for 15 items in the TF test. Calculation of Cronbach alpha internal consistency coefficient was found to be 0.29 for DBT section scores for the matching items in TF. Standard error of measurement was 0.94 for DBT scores and 1.66 for TF test.

Table 2. Test Statistics for DBT and TF Tests Implemented on WG1 (n=159)

Statistics	DBT	TF Test
$\bar{x}$	2.59	10.45
Median	3	11
Mode	3	11
Standard Deviation	0.90	4.36
Minimum	0	5
Maximum	4	15
Average Difficulty	.65	0.70
KR-20	0.03	0.37
Std. Error of Measurement	0.94	1.66

Table 3. Item Statistics of DBT and TF Test Implemented on WG1 (n=159)

DBT			TF Test				
Section	Difficulty	Power of Discrimination	Item No	Difficulty	Power of Discrimination	Average Difficulty	Section Power of Discrimination
1	.62	.63	1	.60	.45	.60	.45
2	.79	.72	2	.67	.53	.77	.56
			3	.87	.38		
3	.64	.43	4	.60	.35	.71	.48
			5	.68	.28		
			6	.64	.12		
			7	.92	.21		
4	.55	.29	8	.80	.36	.69	.76
			9	.65	.34		
			10	.55	.39		
			11	.88	.16		
			12	.32	.27		
			13	.73	.38		
			14	.79	.29		
			15	.76	.24		

Table 3 presents the item statistics regarding DBT and TF tests implemented on WG1. DBT item difficulty index changed between 0.55 and 0.79 whereas item difficulty index for the items in TF test changed between 0.32 and 0.92. Sections starting with section three in DBT were getting harder for the students and the items in TF test that matched with these sections also became more difficult. Power of discrimination indexes calculated for the sections of DBT changed between 0.29 and 0.72 and the values decreased in sections three and four as was the case in item difficulty. Power of discrimination indexes for TF items were found to be between 0.12 and 0.53. Correlation of section total scores in TF test that corresponded with DBT sections with total scores shows that TF had

increased discrimination power compared to DBT starting with section three. Students' midterm scores and their DBT and TF test scores were examined as proof for validity. Midterm exam was composed of 5-option MC items with less chance effect and the scores were obtained with a real exam. The correlations were found to be r=0.17 for DBT and 0.23 for TF. When the effect of other scores were constant, partial correlations were calculated to be r=0.09 for DBT and r=0.17 for TF. The test used in midterm was more comprehensive compared to DBT and TF tests. Examination of 9 MC items, regarding assessment, evaluation, error and scale concepts that are in the framework of DBT, with their scores provides the coefficients of r=.07 and r=.08 respectively.

Response pattern obtained by student answers to the 15 item TF test was examined to identify how the scores may change when items in DBT were placed differently. Places of two items in the second section that assess the identification objectives for assessment types were changed to identify the outputs students would prefer when these items were in different places. New score values were assigned for these outputs and the difference between estimated scores and current DBT scores was investigated with dependent t-test (the results are provided in Table 4). The correlation between DBT scores and estimated DBT scores was found to be 0.51. Scores calculated by changing the places of two items were found to be significantly higher than DBT scores (t=-3.01, p<0.01).

Table 4. Test Results for Comparing DBT Scores and Estimated DBT Scores Obtianed by Changing the Places of Two Items (n=159)

Score Type	$\bar{x}$	S	r	t	p
Obtained from DBT	2.59	.95	.51	-3.013	.003
Estimated	2.82	.97			

Multiple linear regression analysis was undertaken to observe whether DBT scores, TF test scores and estimated DBT scores predicted midterm scores Results show that only the scores obtained from 15-item TF test predicted midterm scores (F=8.35, p<0.01).

Table 5. WG 2 Test Statistics regarding DBT and TF

Statistics	DBT (n=96)	TF Test (n=96)
$\bar{x}$	2.47	2.13
Median	3	2
Mode	3	2
Standard Deviation	0.74	0.91
Lowest	1	0
Highest	4	4
Average Difficulty	0.62	0.53

In the final exam, half of WG2 was assigned DBT whereas the other half was assigned the 4 TF items which were on the route to output with 4 points. Table 5 presents the statistics regarding the scores obtained in the test. As can be seen from the Table, the average scores from DBT were 2.47. Items in the TF test were difficult for students who obtained an average 2.13 scores. The range of scores obtained from TF test was wider and had a higher standard deviation although the average was lower. In other words the scores had a more heterogeneous distribution compared to DBT scores.

Table 6 presents the item statistics regarding DBT and TF implemented on WG2. DBT item difficulty index changed between 0.45 and 0.84, TF test item difficulty index changed between 0.27 and 0.91. Discrimination power index calculated by correlations between DBT item scores and DBT scores changed between 0.17 and 0.65. As was the case in DBT implemented on WG1, values decreased in third and fourth sections as well. Correlations of DBT item scores with final scores were found to be between -0.04 and 0.38. Item difficulty for the TF was calculated to be 0.27 and 0.91 and the average difficulty shows that students found the TF test to be more difficult. Correlation of TF items with the scores obtained from these four items provides values that changed between 0.32 and 0.65. average of these values was higher than the discrimination power average of DBT items (0.09) but examination of the correlations with final scores shows that DBT item correlation to be 0.27 point higher in average.

Table 6. Item Statistics of DBT and TF Test Implemented on WG2

DBT (n=96)				TF Test (n=96)			
Section	Difficulty	Discrimination *	Discrimination **	Item No	Difficulty	Discrimination *	Discrimination **
1	.56	.65	.12	1	.55	.553	-.13
2	.84	.63	.38	2	.91	.32	.08
3	.45	.17	-.04	3	.27	.61	-.34
4	.62	.21	.24	4	.41	.55	.02
Average	.62	.42	.18		.54	.51	-.09

* Correlation with the score obtained from the related section

** Correlation with the scores obtained from the final exam

Relationships between DBT and TF scores of WG2 students and their midterm, final and grade scores with the addition of task scores were examined and results are provided in Table 7. Statistically significant positive correlations were detected between DBT scores and all three types of scores. TF scores showed statistically significant positive correlations with all three types of scores.

Table 7. Validity Coefficients of DBT and TF Tests Implemented on WG2

Test	n	Mid Term	Final	Grade Score
DBT	96	.31**	.40**	.44**
TF	96	-.03	-.20*	-.05

* $p<0.05$, ** $p<0.01$

Following the implementation of DBT, announcement of the scores and the discussions about the 15 TF items, students in WG3 were asked the level of efficiency the tool had about determining their achievement and whether they would use the tool when they became teachers. Student responses to these questions are presented in Table 8. As can be seen from the Table, half of the students in all departments found DBT rather successful in determining their achievements. Students who found the success of the tool to be medium comprised of approximately one third of the groups. There were no students that found the tool to be inefficient. There was only one student who stated that he/she would not use the tool in the future whereas 99.42% of the students stated that they would use DBT when they became teachers.

Table 8. Views of WG3 Students on DBT

Is DBT efficient?	Department									
	Primary Education Math. Teach.		Primary Education Science Teach.		CEIT		Other		Total	
	n	%	n	%	n	%	n	%	n	%
Not At All	-	-	-	-	-	-	-	-	-	-
Not Very Efficient	2	3.0	2	3.1	5	17.2	2	14.3	11	6.3
Moderately Efficient	21	31.8	23	35.4	10	34.5	4	28.6	58	33.3
Rather Efficient	39	59.1	34	52.3	14	48.3	5	35.7	92	52.9
Completely Efficient	4	6.1	6	9.2	-	-	3	21.4	13	7.5
Will use DBT in classroom while teaching?										
Yes	66	100.0	65	100.0	29	100.0	13	92.9	173	99.4
No	-	-	-	-	-	-	1	7.1	1	0.6
Total	66	100.0	65	100.0	29	100.0	14	100.0	173	100.0

Average DBT scores for all levels was investigated to determine whether regarding DBT to be efficient corresponded to DBT scores. Findings are presented in Table 9. Teacher candidates who found the tool completely efficient obtained an average score of 3.50. Students who found the tool to be not so efficient had the lowest average (2.45). Kruskal Wallis analysis shows that DBT scores differentiated according to views on sufficiency of the tool (KWX2= 10.41, $p<0.05$).

Table 9. Relationship between WG3 DBT scores and Finding DBT Efficient

Is DBT efficient?	n	$\bar{x}$	S	Rank Order	X^2	p
Not Very Efficient	11	2.45	1.04	63.64	10.41	0.015
Moderately Efficient	53	2.70	0.97	73.33		
Rather Efficient	83	2.89	0.84	81.62		
Completely Efficient	12	3.50	0.80	113.25		
Total	159*	2.84	0.92			

*Scores for some students could not be obtained since they did not follow the instructions

Discussion, Conclusion and Suggestions

A group of students were given DBT in this study and they were asked to answer to 15 TF items in the DBT as well. Average difficulty was found to be 0.65 following the DBT implementation and it was found to be 0.70 in 15-item TF implementation. In other words, the difficulties of 4-item DBT and 15-item TF are close to each other. The correlation between the scores obtained from both forms was calculated to be 0.45. This medium level correlation between scores obtained from tests that were composed of same items which were answered in both tests and implemented in the same class hour shows that the tests were not highly parallel to each other. In their comparison of DBT and TF, Şeyihoğlu and Erbaş (2010) also found that student achievement changed according to the technique used in the comparison. One reason for low correlation may be based on the fact that TF item type mixes chance effect with scores and has a weak diagnostic value. Another reason may be based on the difference between the two test formats. Internal consistency reliability values of tests show that KR-20 coefficient of the scores obtained from the 15-item test was 0.37 whereas KR-20 coefficient of the scores obtained from DBT was 0.03. Standard error of measurement was calculated to ensure healthier comparisons since variations of item numbers and scores were different. Standard error for scores in 0-4 range in DBT was found to be close to 1. TF test whose score range is twice more than DBT scores and can theoretically provide scores in the 0-15 range provided a standard error of 1.67. These findings show that relatively more errors are observed in DBT scores.

Examination of WG1 DBT and TF test item statistics shows that items became more difficult starting with items answered on the third place in DBT. While percentage of correct answers for the two items in section two was found to be 0.79, the percentage for the items answered the last was found to be 0.55. Examination of items obtained from TF shows that difficulty level was about the same. Starting with Section 3 in which item numbers increased, difficulty in DBT increased while it stayed the same in TF by generating an important discrimination with the TF test which required answers to more questions. Power of item discrimination indexes showed parallels between the first three items and the items in the same section whereas differences in favor of TF were observed in the last section. While the power of discrimination for the last item in DBT was found to be 0.29, this value was found to be 0.76 for the last section of the TF test.

As a proof of validity for the DBT and TF tests, their relationships with the scores of midterm exams in which same content was required and which were composed of MC items were investigated. The results of this investigation pointed to lower validity scores such as r=0.17 and r=0.23. Examination of partial correlations shows almost no relationship between DBT scores and midterm exam scores. When midterm scores are taken as the criteria, it can be claimed that TF has a higher validity compared to DBT although both have low validity scores.

The placement of items in DBT is highly important. Since there are many items in the test and students only use the route pointed by the answers they cannot answer some of the items. Places of the two items in section two were changed in order to see how this would affect their scores and the new scores were estimated based on the student answers previously provided to 15 items. The correlation thus obtained was still medium level (r=0.51). The finding points to the fact that changing the place of two items significantly changes the scores obtained from DBT. Examination of the significance between estimated scores and obtained scores shows that estimated scores were significantly higher than DBT scores ($t=-3.01$, $p<0.01$).

Multiple linear regression analysis undertaken to find whether DBT, estimated DBT and TF test scores predicted midterm test scores pointed to the fact that DBT and estimated DBT scores were not included in the regression equation. In other words, DBT scores calculated in two different manners were not as valid as TF test in which all of the 15 items were implemented.

Along with the final exam, half of WG2 was assigned DBT whereas the other half was assigned the 4 TF items which were on the route to output with 4 points. Average difficulty of these tests implemented to two similar groups shows that DBT was found more difficult by the students. In case that Kocaarslan's (2012) claim that "DBT decreases chance effect compared to TF test" is true, TF test can be expected to be easier in terms of difficulty. Study by Şeyihoğlu and Erbaş (2010) also found lower chance effects for DBT. However the current study found results to the contrary. It was seen that power of discrimination indexes obtained by calculating the relationship between TF test items and the total score were found to be higher compared to DBT sections. When final exam scores were taken as the criteria, it was seen that all coefficients decreased significantly and that discriminatory power of DBT was higher than those of TF test items. Relationships between DBT and TF scores and students' midterm, final and final grade scores show that DBT had a higher validity coefficient. This finding can be interpreted that DBT is more valid than a 4-item TF test.

Students in WG3 were asked how efficient they found DBT in determining their achievements following the implementation of DBT and announcement of their scores. An interesting finding is the fact that none of the teacher candidates found this tool to be inefficient. The ratio of the students regarding the tool to be not so efficient was found to be 6.3%. More than half of the students found the tool to be highly efficient or completely efficient to determine their achievements. Examination of the relationship between finding DBT to be efficient and the score obtained from DBT shows that students who found DBT efficient were the students who obtained high scores from DBT and that students with low scores regarded the tool as inefficient.

Studies on DBT point that teachers do not generally use DBT in their classroom since they do not regard themselves competent or informed enough to use DBT, that it takes a long time to prepare and that classrooms are too crowded to use the tool (Aydoğmuş and Coşkun Keskin, 2012; Çepni and Şenel Çoruhlu, 2010;; Karamustafaoğlu, Çağlak and Meşeci, 2012; Okur, 2008; Özdemir, 2010; Pullu, 2008; Şaşmaz Ören, Ormancı and Evrekli, 2011; Şenel Çoruhlu, Er Nas and Çepni, 2009). In the current study, almost all teacher candidates (99.4%) reported that they would use DBT in their classrooms when they became teachers.

Results regarding the implementation of DBT to three different groups are as follows:

- DBT composed of true-false items includes the disadvantages of this item type. Chance error is very high and its diagnostic power is weak. The fact that the name of the tool includes the term "diagnostic" cannot go beyond formality.
- High ratio of chance effect which is a type of random error negatively affects the reliability and validity of the scores obtained with this tool. More valid scores were obtained with the use of DBT compared to 4-item TF test but when all 15 items in the TF test are implemented, the scores will be more valid and reliable.
- Teacher candidates regard DBT as a valid tool. While active teachers do not use the tool for various reasons, teacher candidates who were given DBT implementations stated that they would use the tool in their classes.
- Findings were obtained in studies on DBT that the use of DBT supports teaching but any tool has to be valid, reliable and useful in order to be beneficial. Findings obtained in this study show that efficiency of DBT is limited in these properties. It is also evident that it cannot assess any properties that cannot be assessed with traditional tools. Therefore, it can be claimed that it may not be appropriate to include DBT among performance based assessment methods.

References

Aydoğmuş, A. & Coşkun Keskin, S. (2012). Use of process based measurement and evaluation tools by social sciences teachers: İstanbul province sample. *Mersin University Faculty of Education Journal*, 8 (2), 110-123.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, 21 (3), 5-18.

Çepni, S. & Şenel Çoruhlu, T. (2010). Reflections on teaching from in-service training course based on alternative assessment and evaluation techniques. *Pamukkale University Faculty of Education Journal*, 28 (1), 117-128.

Çıkrıkçı-Demirtaşlı, N. (2012). Relationships among learning, teaching and evaluation, N.Çıkrıkçı-Demirtaşlı (Ed.) *Measurement and Evaluation in Education* (pp. 3-29). Ankara: Elhan Publications.

Gelbal, S. & Kelecioğlu, H. (2007). Teachers' perceptions of competence regarding measurement and evaluation methods and the problems they faced. *Hacettepe University Faculty of Education Journal*, 33, 135-145.

Johnstone, A. H., McAlpine, E. & MacGuire, P.R.P. (1986). Branching trees and diagnostic testing. *A Journal for Further and Higher Education in Scotland*, 2, 4-7.

Karahan, U. (2007). *Adoption of grid, diagnostic branched tree and concept maps; alternative assessment and evaluation methods; to teaching biology*, Unpublished Master's Thesis, Gazi University Educational Sciences Institute, Ankara.

Karamustafaoğlu, S., Çağlak, A. & Meşeci, B. (2012). Classroom teachers' self-competencies related to alternative assessment and evaluation tools. *Amasya University Faculty of Education Journal*, 1 (2), 167-179.

Karasar, N. (2000). *Scientific research method (10. Edition)*. Ankara: Nobel Publication and Distribution.

Kocaarslan, M. (2012). *Diagnostic branched tree technique and its use in 5th grade science and technology class unit "change of matter and recognition"*. Mustafa Kemal University Social Sciences Institute Journal, 9 (18), 269-279.

Kutlu, Ö., Doğan, C.D. &Karakaya, İ. (2008). *Identification of student achievement based on performance and portfolio*. Ankara: Pegem Academy.

Linn, R. L. & Gronlund, N.E. (1995). *Measurement assessment in teaching*. New Jersey: Prentice Hall Inc.

MoNE (2013). *Curriculums*, Turkish Education Board, http://ttkb.meb.gov.tr/www/ogretim-programlari/icerik/72 , accessed on 19.03.2013.

MoNE (2005). *Primary education 1–5. grades* curriculums *information booklet*. MoNE, TEB Directorate, Directorate of Education-Training and Curriculums Unit. Ankara: State Books Directorate Publication.

Okur, M. (2008*). Identification of 4. and 5. Grade classroom teachers' views on alternative assessment and evaluation techniques in science and technology classes*. Unpublished Master's Thesis, Bülent Ecevit University Social Sciences Institute, Zonguldak.

Özdemir, S. M. (2010). Primary school teachers' competences regarding alternative assessment and evaluation tools and their in-service training needs. *Turkish Educational Sciences Journal*. 8(4), 787-816.

Pullu, S. (2008). *Classroom teachers' views on assessment and evaluation in primary school curriculums and their practices (elazığ province sample).* Unpublished Master's Thesis, Fırat University Social Sciences Institute, Elazığ.

Şaşmaz Ören, F., Ormancı, Ü. & Evrekli, E. (2011). Self-competence levels and views of science and technology teacher candidates in alternative assessment and evaluation approaches. *Educational Sciences in Theory and Practice*, 11(3), 1675-1698.

Şenel Çoruhlu, T., Er Nas, S. & Çepni, S. (2009). Problems faced by science and technology teachers in using alternative assessment and evaluation techniques: trabzon sample. *Yüzüncü Yıl University Faculty of Education Journal*, 1(I), 122-141.

Şeyihoğlu, A. & Erbaş, A.A. (20-22 May 2010). *Comparison of diagnostic branched tree technique and true-false technique in social sciences lesson*. 9. National Classroom Teaching Symposium, Elazığ.

Şimşek, N. (2004). A critical approach to constructivist teaching and learning. *Educational Sciences and Practice*, 3 (5), 115-139.

TLA (2013). *Grand turkish dictionary*. http://tdkterim.gov.tr/bts/ accessed in 19.03.2013.

Tekin, H. (1996). *Assessment and evaluation in education (9. Edition),* Ankara: Yargı Publications.