Participatory Educational Research (PER)
Special Issue 2016-III, pp., 68-76 November, 2016
Available online at http://www.partedres.com
ISSN: 2148-6123

Development and Validation of the Achievement Test on Body Systems

Aslı YERLİKAYA*and M. Handan GÜNEŞ
Department of Science Education, Ondokuz Mayıs University, Samsun, Turkey

Abstract
This paper presents evidence for the development and validation of an "Achievement Test on Body Systems" which is a unit in 7th class Science Education lesson. The test was conforming the acquisitions described in the Turkish National Science Education Program for 3, 4, 5, 6, 7 and 8th classes. At the same time, number of the acquisitions, step of the acquisitions according to Bloom's taxonomy, number of lessons to finish all units and formerly learned concepts were taken into account for creating items. The content validity of the test was reviewed and ensured by 1 expert and 2 science teachers. One of the teachers was doing her PhD degree and one of them was doing his master's degree. The pilot test was made up of 48 items selected from initial 78 items and the sample consisted of 245 7th grade students. Kaiser-Meyer-Olkin Test (which is a measure of how suitable the data are for Factor Analysis) was done. Kaiser-Meyer-Olkin value was 0.89, and the result indicated that the sampling was adequate. Analyses based on classical test theory used 33% of the upper and lower classes for item difficulty and discrimination indices. And the KR-20 was used for the reliability. After omitting 3 items; the item difficulties were between 0.36 and 0.81, the item discrimination indices were between 0.40 and 0.89 and the internal consistency of the test (KR-20) was found as 0.898. Results indicate that Achievement Test of Body Systems has satisfactory psychometric properties and it can be accepted as appropriate for future use.

Key words: achievement test, validity and reliability, body systems

Introduction

Various philosophies, approaches, methods and techniques are used when the learning-teaching processes are being planned, developed and applied. The definition of the knowledge by each philosophical movement is different from each other (Sonmez, 2009). In this context, approaches that are based on certain philosophical principles and that develop based on these philosophical bases, and the methods and techniques, which are the applications of these approaches in the education medium, vary in a great deal (Durmus, 2005; Demir, 2009; Bezir Akcay, 2014). In order to understand whether various methods and techniques used in educational media are beneficial in teaching, and to learn how much

* yerlikayasli@gmail.com

students acquire the subjects and concepts given to them, it is necessary to make measurements and evaluations (Izard, 1993; Mertens, 2015).

Making evaluations on whether the works done in each field of life within a certain program achieve success or not, or on determining how much they achieve success are extremely important, and the criterion of these evaluations reveal the level of success. As it is already known, curricula include elements, which are, the target, content, application-learning experiences and evaluation. When considered in this context, it attracts attention that evaluation is an inseparable part of education. As a matter of fact, teachers have to find the most suitable measurement and evaluation technique to see how much the program is effective or to determine what students learn during the educational processes (Balaban and Gunes, 2012).

Measurement is defined as observing a measurable quality and expressing this quality with numbers and symbols that are proper for the purpose. The most basic problem in this process may be experienced in determining a unit system that will facilitate the performing of the measurement process with equal units, and that is generalizable and suitable for the purpose. In education, on the other hand, measurement is described as determining how much the behaviors change in the direction of the targeted purposes, and showing this with various techniques, numbers and symbols (Cepni et al., 2007).

Evaluation is defined as the determination of whether the expected behaviors are acquired by students in the education process or how much these behaviors are acquired (Atilgan et al., 2006).

These two concepts, which are confused with each other, are generally given together. However, measurement constitutes one of the most basic elements of evaluation concept. For example, determining how many points a student receives from a test is measurement, and determining whether that students passes this class with these points or not is evaluation (Demirel, 2002).

A measurement tool that has to be reliable, valid and functional may be applied before, during and/or after the teaching activity (Costu, 2012) in various forms. These forms may be in the form of written examination, oral examination, multiple-choice tests, tests with short answers, true-false tests, etc. Of course, when these tests are being prepared, there are some points that have to be cared for, such as the questions must not be ambiguous, they must be in accordance with the acquisitions, there must not be information having clues for the answers in the questions or in the options, the test must be examined by other people not only by the one who prepares them, possible mistakes must be corrected, and the difficulty level must be adjusted well (Gecit, 2013).

By considering all these points, the aim is to develop a multiple-choice achievement test on the unit “The Systems in Our Body”, which is taught in 7th Grades. In this context, firstly, the issue of how the subject is given in the curriculum has been examined. When considered in general terms, the information on the systems of our body, which we use every day, is given in 3, 4, 5, 6, 7 and 8th grades in the scope of Science Education Classes in primary schools (Ministry of National Education, 2013). In 7th grades, on the other hand, there are the parts named “Digestive System”, “Excretory System”, “Auditory and Regulatory Systems”, “Sense Organs” and “Organ Donation and Transplantations of Organs” in the unit called “The Systems in Our Body”. In the curriculum about this unit, there are 16 acquisitions

in the 7th Grade and 4 of these acquisitions are given in the subject of "Digestive System" and 2 are given in "Excretory System", 4 in "Auditory and Regulatory Systems", 5 in "Sense Organs", and 1 in "Organ Donation and Transplantations of Organs". When the subject is assessed by considering the acquisitions given in the previous years and in the following year in other levels to make associations by students; it is observed that there are 3 acquisitions in the 3rd Grade, 8 in the 4th Grade, 13 in the 5th Grade, 14 in the 6th Grade, and 13 in the 8th Grade (Ministry of National Education, 2013).

Methodology

In order to develop an achievement test on "The Systems in Our Body" unit in 7th Grade, first of all, the purpose and the concepts-acquisitions to be measured were determined by examining the curriculum released by the Ministry of National Education, Board of Education and Discipline (2013). Then, the items were written, the viewpoints of specialists were referred to, the possible format of the test was prepared and applied. The application results of the test, item analyses, and item selection were made, and the last form of the test was given by considering the statistical results.

Sample

The sampling of the study consisted of 245 seventh grade students who were studying at state schools in Samsun in 2015-2016 Academic Year. In order to reveal the real situation in the universe, the most important condition is the sampling to represent the universe. For this, the number of the sampling must be adequate. With this viewpoint, in order to determine whether the sampling was adequate or not, the Kaiser-Meyer-Olkin (KMO) Test (sampling adequacy criterion) was performed (Kaiser, 1970).

Item Pool (Content Validity)

As the first step or as the developing the achievement test step, an item pool consisting of 78 items was formed. When forming this pool, the curriculum that was run by the Ministry of National Education, Board of Education and Discipline (2013) was examined. In these examinations, the information on the subject learnt in the previous year, the number of the acquisitions in 7th Grade in "The Systems in Our Body" unit, its place in the unit as a curriculum item, and the corresponding elements of these acquisitions in Bloom Taxonomy were considered. By doing so, the purpose was to ensure the content validity. Validity is related with whether a study design measures the subject to be measured (McCowan & McCowan, 1999; Erdogan, 2012), in other words, it is related with how truly the test measures the characteristics of an individual (Demir, Gurer, Koksal & Dolu, 2009; Buyukozturk, 2011). After the items of the measurement tool and its structure were produced, the item analysis and the reliability and validity constitute the indirect evidence on the content validity of the measurement tool (Erkus, 2011).

The Viewpoints of the Specialists (Content Validity)

In order to ensure the content validity of the achievement test, 78 items were prepared at first. Specialist viewpoints were referred to in order to examine whether each item had the

quality of measuring the desired skill, in order to examine the mistakes in terms of spelling and meaning, the items not giving one another's answer, and in order to determine whether they are true in terms of science. For this purpose, 78 items were examined by 2 Science Education Teachers one of whom was studying for doctorate degree and one of whom continued post-graduate degree, 1 Turkish Language teacher, 1 measurement-evaluation specialist, and field specialist. In the light of the specialist viewpoints, the draft items underwent a pre-selection process, and the test was reduced to 48 items. The pilot study of the achievement test was applied to the sampling with 48 items.

Reliability Studies with Factor Analysis

Reliability studies with factor analysis were conducted with computer programs (SPSS and Excel). Firstly, the students were encoded like S1, S2. Then, the answers given by students for each item (which option was chosen by the student) was saved in an Excel file. After this step, the value of 1 was given to the option that was true according to the key of the test, and 0 was given to the wrong answers. By using the data obtained, the KMO, Distinguishing Power Index of the Items (r), Item Difficulty Index (p), and the Kuder-Richardson 20 (KR-20) values were computed.

Results

The data obtained in the analyses are as follows under these titles, respectively;

- The Findings on Evaluating the Conformity of the Data with Factor Analysis,
- Findings on Distinguishing Power Index of the Items and Item Difficulty Index,
- Findings on Validity and Reliability,
- Findings on the Distribution of the Points of the Test Developed.

The Findings on Evaluating the Conformity of the Data with Factor Analysis (the KMO Test)

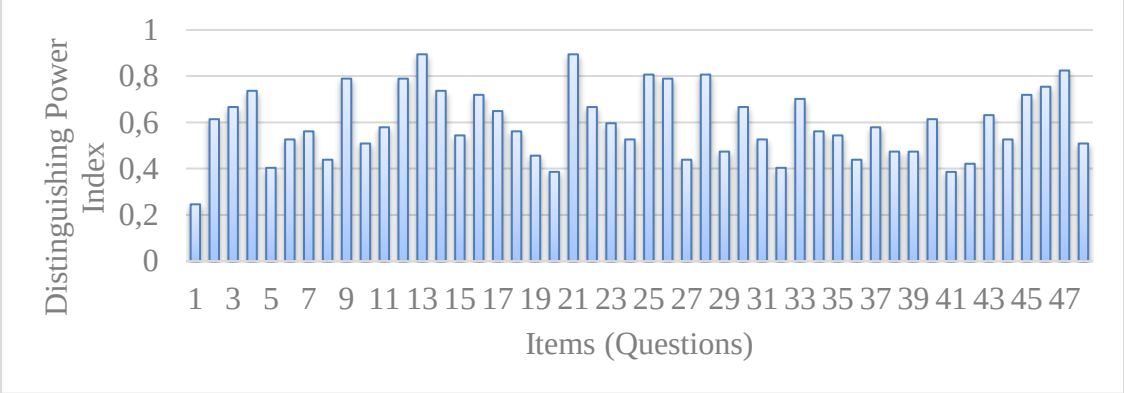
The issues of whether it is correct to perform factor analysis with a data group; and if it is correct, how much it is suitable for analysis are determined by using Kaiser-Meyer-Olkin Test (Kaiser, 1970, 1974). As it is emphasized by Kaiser (1970), this test is not a test statistics, but is a criterion. According to Kaiser (1974), the results of this criterion receive values between 0 and 1. The ones with 0,90 value are interpreted as being "Perfect"; 0,80 as "Good"; 0,70 as "Medium Level"; 0,60 as "Weak"; 0,50 as "Bad" and below 0,50 as "Unacceptable". The value of this criterion was obtained by using SPSS, and was determined as 0,891. When the range of the criterion was examined, it is observed that the achievement test, which was developed, was between the value ranges that are interpreted as "Good". In other words, the sampling in the study represents the universe, and therefore, it was concluded that accurate analyses could be made.

Findings on Distinguishing Power Index (r) of the Items and Item Difficulty Index (p)

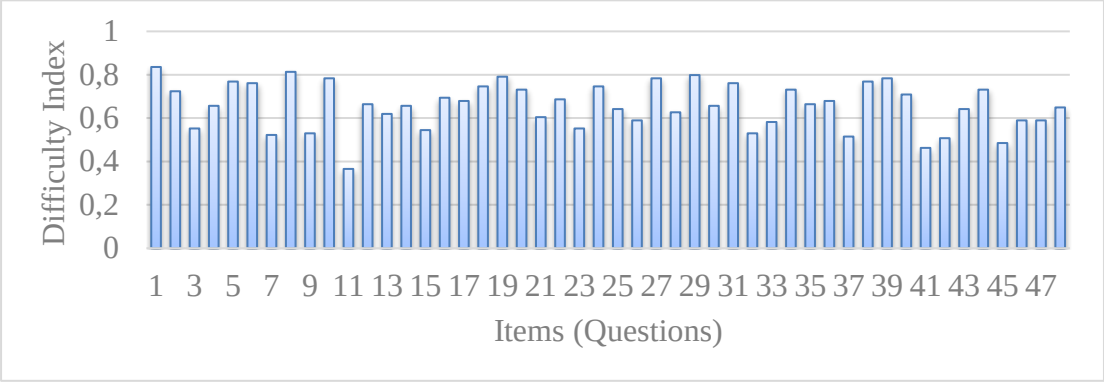
Distinguishing power index of the items (r) is the distinguishing of an item between a knowing and unknowing student (Brennan, 1972; Mayo et al., 1993). In other words, the higher distinguishing power index of an item, the higher it distinguishes between knowing and unknowing students (Masters, 1988). Item difficulty index (p), on the other hand, is the determination of the difficulty level of an item (Ding et al., 2006). The r takes values between -1 and 1; and the p value takes values between 0 and 1 (McCowan & McCowan, 1999). The grouping made by Ebel (1972) has been used by many researchers in their studies (for example, McCowan & McCowan, 1999; Ding et al., 2006). If the r value of the item is or over 0,40 it is interpreted as "Very Good"; between 0,30 and 0,39 as "Good"; between 0,20 and 0,29 as "It must be corrected"; between 0 and 0,19 as "It must be excluded". If the item takes 0 or a negative value, it is understood that the item is written in a bad way, and although it may be answered easily by the students from the lower groups, it is not answered by the students from the upper group. If the p value, on the other hand, is or below 0,40 it is interpreted as "Difficult"; higher than 0,40 and lower than 0,60 "Medium"; and equals or is higher than 0,60 "Easy". In other words, as the value comes closer to 1, the number of the students who know the answer to the question increases, and therefore, the question becomes easier. As the item difficulty index comes closer to 0, the number of the students who know the answer to the question decreases, and therefore, the question becomes difficult.

According to Kelley (1939), in order to make an item discrimination power analysis, taking 27% of the "extreme" groups will be "ideal". According to Wiersma and Jurs (1990), it becomes easier to reveal these differences because the differences increase by moving away from normal distribution (as cited in McCowan and McCowan, 1999). For this reason, the data are ranked from the students that receive the highest points to the ones that receive lowest points. The lower and upper groups (67 students each) were determined in the light of this information.

Graphic 1. The Distinguishing Power Index Values (r) of the Items in the Test.



Graphic 2. The Item Difficulty Index Values (p) of the Items in the test.



The p and r values determined for all items are given in Graphic 1 and 2. As a result of evaluating these two values together, it was determined that the discrimination power of 1st, 20th and 41st items was found to be low, and were excluded from the test. The number of the questions in the achievement test decreased to 45 after some items were excluded. While the average p value of the achievement test was 0,656; the average r value was determined as 0,612. In other words, the Achievement Test consisting of 45 questions consisted of "easy" and "very good" items.

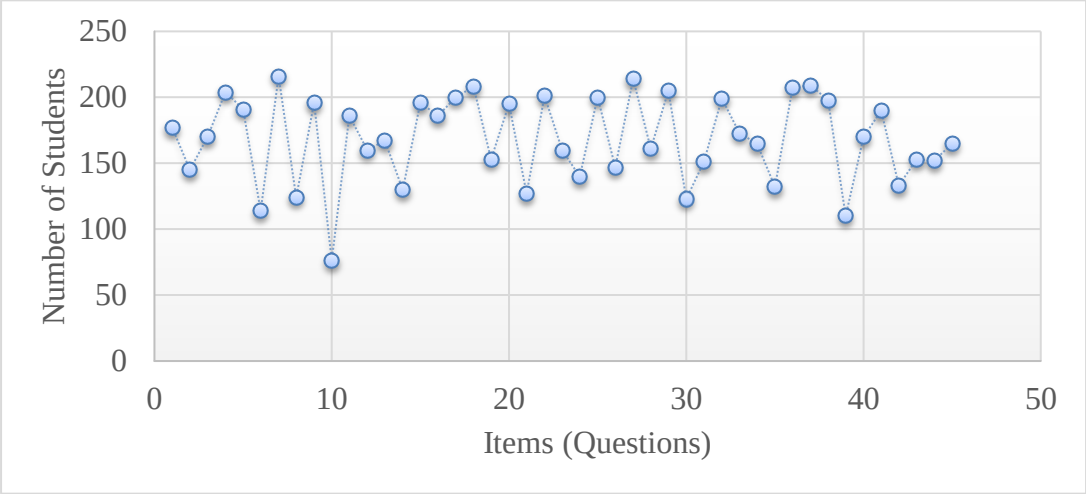
Findings on Reliability (Kuder-Richardson 20)

Reliability is related with whether the measurement will reveal consistent and balanced results or not (Erdogan, 2012). Turgut (1990) defined reliability as the criterion of clearing the measurement results from random errors; Crocker and Algina (1986) defined it as the repeatability of the measurement results that are conducted for the purpose of measuring a certain attribute on the same individuals in similar conditions (as cited in Buyukozturk et al., 2010). When the reliability of the achievement tests is determined, the KR-20 value is considered (Karasar, 2010). KR-20 is a method based on one single application for the purpose of determining the internal consistency reliability of the achievement tests (Erkus, 2011). KR-20 received values between 0 and 1, and the 0,70 value is an acceptable value. As a result of the analyses conducted on the achievement test consisting of 45 questions, the KR-20 value was evaluated, and this value was determined as 0,898. It was concluded that the developed scale was reliable.

Findings on Distribution of the Points of the Test Developed

In addition, if 1 point was given to each question in the achievement test, which consists of 45 questions, the average point obtained from the data was determined as 31,09. While the average value of the upper group students who answered the questions correctly was 61,3 (91,5%); this average was 26,4 (39,4%) in the lower group. The definitive statistical findings are given in Graphic 3.

Graphic 3. The Frequency of the Students in Giving Accurate Answers Based on Questions



Discussion

In this study, the purpose was to develop an achievement test to measure the success of the students on the Body Systems unit, and the test was developed by applying and performing its validity-reliability and item analyses. In this context, the KMO criterion, validity, reliability and item analyses were performed for the "Achievement Test on Body Systems", which consisted of 45 items; and each question had 4 options at 7th Grade level. All of these results are given altogether in Table 1.

When the item difficulty index value of all the items in the achievement test were examined, it was observed that the values of the items varied between 0,37 and 0,81. It appeared that the test was an easy achievement test in terms of the difficulty level. The discrimination power index values of the items received values between 0,40 and 0,89. The discrimination power of the items is extremely high in terms of discrimination power. For this reason, it has been concluded that an Achievement Test with items, which discriminates between the knowing students and those not knowing and which has mostly easy items, has been developed. In addition, it has been revealed that this test, whose KR-20 value is 0,898 in terms of internal consistency reliability, is a reliable achievement test. When the scale was being developed, meanwhile a valid achievement test was developed with the help of sticking to the concepts and acquisitions and the specialist viewpoints and item pool. In other words, an achievement test has been formed, which discriminates between the knowing students and those who do not know, and which is reliable and valid in this aspect. By doing so, an achievement test, which is suitable for the acquisitions and purposes, analyzed by using adequate sampling, and which has adequate validity and reliability has been developed. In other words, it may be claimed that the "Achievement Test on Body Systems" at 7th Grade level is a test that can measure the knowledge of the students on this topic in an accurate manner.

Table 1. The Statistics and Criterion Found in This Research

Test Statistics / Criterion (KMO)	Between	Acceptable Levels	Achievement Test's Items Levels	Achievement Test's Items Assessment
KMO	0 / 1	≤ 0,50	0,891	"good"
p	0 / 1	≤ 0,30	0,656	"easy"
r	-1 / 1	≤ 0,30	0,612	"very good"
KR-20	0 / 1	≤ 0,70	0,898	"reliable"

For this reason, future researchers, who aim to investigate this topic in the future, can use "Achievement Test on Body Systems" in order to measure the influence of various methods and techniques on learning.

In addition, it is also considered that this study will cast a light for the researchers who would like to develop achievement tests to be applied on different topics or in different educational levels. For this reason, the development steps for the achievement test have been explained in detail.

References

Atılgan, T. L., Doğan, N., Kan, A., (2006). *Eğitimde Ölçme ve Değerlendirme.* (Editör: Hakan Atılgan). Anı Yayıncılık, 480 s. Ankara.

Balaban, M. and Güneş, M. H. (2012). Portfolio assessment in cooperation with teachers, students and parents in a science and technology course. Eurasian Journal of Educational Research, Issue 49/A, 289-310.

Bezir Akçay, B. (2014). Bilimde Paradigmalar ve Bilimin Doğası. Şengül S. Anagün ve Nil Duban (Ed.) içinde Fen bilimleri Öğretimi. Ankara: Anı Yayıncılık.

Brennan, R. L. (1972). A Generalized Upper-Lower Term Discrimination Index. Educational and Psychological Measurement, 32 (2), 289-303.

Büyüköztürk, Ş. (2011). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem Yayıncılık.

Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., Karadeniz, Ş. & Demirel, F. (2010). *Bilimsel araştırma yöntemleri.* Ankara: Pegem Yayıncılık.

Çepni,S., Ayas, A., Akdeniz, A. R., Yiğit, N., Özmen, H., Ayvacı, H. Ş. (2007) *Kuramda Uygulamaya Fen ve Teknoloji Öğretimi.*(Editör: Salih Çepni), Pegem A. Yayıncılık, 431 s., Ankara.

Coştu, B. (2012). Ölçme ve Değerlendirmeyle İlgili Temel Kavramlar. Mehmet Küçük ve Yılmaz Geçit (Ed.) içinde Eğitimde Ölçme ve Değerlendirme, Ankara: Nobel Yayıncılık.

Demir, M. Y. (2009). Öğrenme: Giriş, Sorunlar ve Tarihsel Bakış Açıları. Muzaffer Şahin (Çev. Ed.). içinde Öğrenme Teorileri. Ankara: Nobel Yayın Dağıtım.

Demir, S., Gürer, C.,Köksal, T. & Dolu, O. (2009). *Kavram Oluşturma ve Ölçüm*. *(Ed.) Kaan Böke. Sosyal Bilimlerde Araştırma Yöntemleri.* İstanbul: Alfa Yayıncılık.

Demirel, Ö. (2002). *Planlamadan Değerlendirmeye Öğretme Sanatı*. Pegem A Yayıncılık, 431. s., Ankara.

Ding, L., Chabay, R., Sherwood, B. & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. Physical Review Special Topics - Physics Education Research 2, 010105, 2 (1),

Durmuş, S. (2005). Öğrenme: Perspektifler. Anakara: Nobel Yayın Dağıtım.

Erdoğan, İ. (2012). Pozitivist Metodoloji ve Ötesi. Ankara:ERK Yayınları.

Erkuş, A. (2011). *Davranış Bilimleri İçin Bilimsel Araştırma Süreci*. Ankara: Seçkin Yayıncılık.

Geçit, Y. (22012). Geleneksel Ölçme Araçları ve Özellikleri. Mehmet Küçük & Yılmaz Geçit (Eds.) içinde Eğitimde Ölçme ve Değerlendirme.

Izard, J. (1993). Challenges to the Improvement of Assessment Practice, In M. Niss (Ed.). Investigations Into Assessment in Mathemetics Education, Kluver Academic Publishers, Netherlands.

Kaiser, H. F. (1970). A second generation little jiffy. Pschometrika, 35 (4), 401-415.

Kaiser, H. F. (1974). An index of factor simplicity. Pschometrika, 39 (1), 31-36.

Karasar, N. (2010). *Bilimsel Araştırma Yöntemi.* Ankara: Pegem Yayıncılık.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30 (1), 17-24.

Masters, G. N. (1988). Item Discrimination: When More Is Worse. Journal of Educational Measurement, 25 (1), 15-29.

Mayo, P., Donnelly, M. B., Nash, P. P., & Schwartz, R. W. (1993). Student Perceptions of Tutor Effectiveness in a Problem-Based Surgery Clerkship. Teaching and Learning in Medicine, 5 (4), 227-233.

Mertens, Donna M. (2015). Research and Evaluation in Education and Psychology: Integrating Diversity (4th edition). Sage Publications.

McCowan, R. J. & McCowan, S. C. (1999). Item Analysis for Criterion-Referenced Tests.

Milli Eğitim Bakanlığı [MEB] (2013). İlköğretim Kurumları Fen Bilimleri Dersi (3, 4, 5, 6, 7 ve 8. Sınıflar) Öğretim Programı, Ankara: Talim Terbiye Kurulu.

Sönmez, V. (2009). Eğitim Felsefesi. Ankara: Anı Yayıncılık.