WCES 2014

A Study For Developing A Success Test: Examination Of Validity And Classification Accuracy By ROC Analysis

Fazilet Taşdemir[a] *

[a] Faculty of Educational, Recep Tayyip Erdoğan University, Rize, 53200,Turkey

Abstract

In this research the aim is to develop a success test including grammar, reading-comprehension, and vocabulary and sentence knowledge in order to determine the Turkish lesson learning skills of students at secondary school fourth grade. In determination of cognitive property stages the comprehension processes of PIRLS were used. While the research group was determined a suitable sampling method among the non-random sampling methods was used. The research group consists of 111 students having education in secondary school fourth grade in Etimesgut province of Ankara. Open ended questions were developed as well as multiple choice questions and rubrics were developed for open ended questions. In success test validation study, ROC analysis was performed. Material and test statistics were calculated on the data obtained at the end of the study. The KR20 reliability coefficient was calculated as 0.76 for Turkish success test. With material analysis the average material difficulty and distinctiveness indexes were calculated respectively as 0.49 and 0.34. These results show that valid and reliable results are obtained by the developed test. At the end of ROC analysis, the area under the operation characteristic curve (AUC) was calculated as 0.88 and was found to be significant. (p<0.05) This value shows that Turkish success test can make a separation in determining the student success which means successful and unsuccessful students can be classified with an accuracy ratio of 88% (good level). This result also shows that ROC analysis can be used as a validity detection tool in test development studies.

1. Introduction

Achievement, in general terms, refers to a progress towards a desired result (Wolman, 1973). Achievement in

*Fazilet Taşdemir. Tel.: +0-505-662-1389; fax: +9-0464 532 8612
  E-mail address:fazilettasdemir@mynet.com; fazilet.tasdemir@erdogan.edu.tr

doi:10.1016/j.sbspro.2015.04.473

education, on the other hand, refers to expression of skills and acquired knowledge determined by marks developed in lessons and given by teachers, test scores or both as "Academic Achievement" (Carter and Good, 1973). Achievement tests are prepared and used to determine the progress the students make in terms of knowledge, concept and understanding at the end of a teaching period which is based on a certain program (Yıldırım, 1999).The process of developing achievement tests is given more importance than other tests as they are more frequently applied than other psychological measurement tools. If used effectively, the results of achievement tests serve as a tool to evaluate and motivate students; it also provides information on the objectives of the education to teachers, school administrators and parents. Obtaining data on the effectiveness of education program and teaching processes, test scores contribute to planning of single education processes for classes and students (Aiken, 1994). A major part of learning-teaching attempts at schools aim acquiring cognitive skills by students (Özçelik, 1998). With the Progress in International Reading Literacy Study (PIRLS) of The International Association for the Evaluation of Educational Achievement (IEA), the issues such as reading skills of students, reading habits, teaching methods used by teachers to make students acquire reading skills, whether the teaching materials are sufficiency and the contributions of families to acquisition of reading skills by the students are determined by international standard test and questionnaires and similarity and differences are determined by comparing the data of the participating countries (EARGED, PIRLS 2001 National Final Report, 2003). PIRLS concentrates on three aspects of reading which are reading aims, reading habits and attitudes towards reading. Comprehension processes are as follows: (Mullis, Martin, Kenndey, Trong and Sainsbury, 2011; EARGED, PIRLS 2001 National Final Report, 2003):
Comprehension Processes;
Level 1: Focusing on explicityly explained knowledge and ideas, making implications,
Level 2: Making direct implications,
Level 3: Combination and interpretation of ideas and knowledge,
Level 4: Analysing and evaluating the components, content and language of the text.

This study tested the validity of the test using ROC analysis apart from validity studies used in achievement test development studies. Aydemir et al., (2006) adapted Borderline Personality Inventory into Turkish culture and tested its reliability and validity. Validity analysis examined specificity and sensitive; cutting point was obtained by ROC analysis and in this cutting point, sensitivity was identified as specificity. As a result, it was concluded that the inventory could successfully distinguish personality disorder group from other patient groups and healthy individuals and that it can be used as a reliable and valid tool. Başgül, Etiler, Coşkun, Karakaya and Ağaoğlu (2009) aimed to examine the validity and reliability of Early Childhood Invesntıry-4 parents form (EÇE-4:EF) which evaluates psychological disorders according to diagnostic classification of The Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). The researchers analysed the validity of inventory using sensitivity and specificity criteria. Furthermore, using ROC curves, they determined cutting points of the scale. They found that Early Childhood Inventory was valid and reliable to evaluate common psychological disorders in Turkish children.

This study aimed to develop a Turkish achievement test containing PIRLS comprehension processes to measure the learning of secondary school fourth grade students in "semantics/comprehension" unit and to examine the validity of the test and the accuracy in classification of students in sacksful/unsuccessful groups.

2. Method

This is an achievement test development study. This section includes information of study group and data collection.

2.1. Participants

Among non-random sampling methods, appropriate sampling method, which is a purposeful sampling method, was used to determine study group. Study group consisted of a total of 111 students enrolled in secondary school fourth grade in Etimesgut district of Ankara.

2.2. Procedure

An achievement test containing grammar, reading-comprehension, vocabulary and syntax was developed to determine learning level of the students. A table of specifications was prepared about the unit and a test consisting of comprehension processes including semantics (50%), vocabulary (20%), syntax (20%) and grammar (10%) was applied to measure the learning of the students within the scope of Turkish lesson. The students were given 25 minutes to complete 12 multiple choice items. Attention was given to the significance levels of the determined criteria and mental skill dimensions and by designing a table of specifications accordingly, number of test items for levels and their percentages were calculated. While testing behaviors, whether the semantics of Turkish words and sentences, grammar rules were learned was analyzed. Items were developed at levels in PIRLS.

3. Statistical Analysis

Findings obtained by statistical analyses are listed below.

1. Item and test statistics of the developed test were calculated. KR-20 reliability coefficient for Turkish achievement test was found to be 0.58. This result indicated that the test has a moderate level of internal consistence. ($\bar{X}$=9,8, mod=10, median=10) were calculated as test statistics. These results showed that the test was normally distributed. According to the item analysis, item difficulty and discrimination index varied between (0.49-0.88) and (0.34-0.56) respectively. Based on item difficulty and item discrimination indexes it was found that the items were at acceptable levels and can be used in the test.

2. ROC analysis was used in the study to determine the validity and classification accuracy of Turkish achievement test. Reference points to be used for ROC analysis included Turkish lesson marks of the students. ROC enables determining appropriate cutting points to identify optimal sensitivity and optimal specificity of a test. ROC curve is a method used to determine the performances of individuals in tests and to evaluate the accuracy of statistical model such as linear classification analysis (Obuchowski, Lieber and Wians, 2004). AUC determines accuracy rate of the test in discriminating successful and unsuccessful people. Larger area below ROC curve shows better estimates of the test in accurate classification (Grove, 2006; Hanley and McNeil, 1982). Normality of the distribution of the groups is analyzed before ROC analysis. Kolmogorov-Smirnov statistics were found to be between (0.27 and 0.31) in successful and unsuccessful groups (p> 0.01). Close values in statistics such as mean, median and mod, indicated that the assumption of normal distribution in groups was satisfied.

ROC analysis was conducted taking Turkish lesson achievement mean marks of the students to determine actual conditions of the students who were classified as successful/unsuccessful according to Turkish test scores. A certain cutting point was taken as a basis to determine cutting points that accurately discriminates successful /unsuccessful students. The students in the range of (55 and 69/ mod) scores according to their grade point averages were separated to determine appropriate cutting points and the ROC curve was formed. ROC analysis showed that (AUC)=0.88 and was significant (p<0.05). This value indicated that Turkish achievement test could make discrimination in determining achievement of students. In other words, the test could classify the students with a ratio of 88% (a good level). To decide which value should be taken as the cutting point, each sensitivity and 1-specicity value given at the end of the analysis was examined and optimum point was selected.
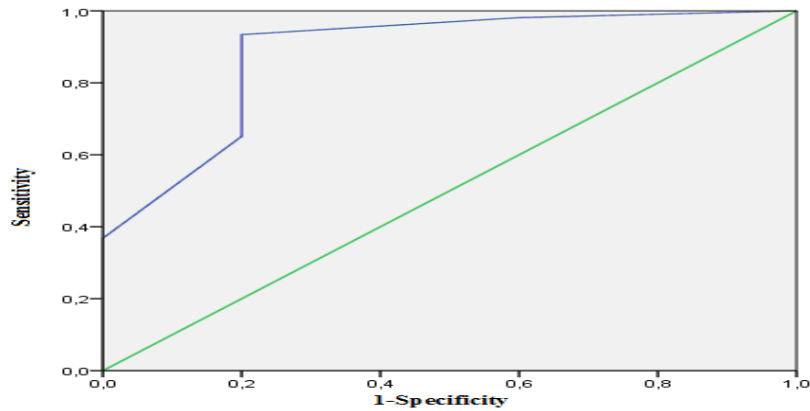
Figure 1.ROC curve

As indicated in Figure 1, this cutting point first makes a peak then a plateau. The coordinates for cutting points determined by ROC analysis are presented in Table 1.

Table 1. Coordinates of the curve

Cutting point	Sensitivity	1-Specificity
5,00	1,000	1,000
6,50	,981	,600
7,50	,934	,200
8,50	,840	,200
9,50	,651	,200
10,50	,368	,000
11,50	,160	,000
13,00	,000	,000

According to Table 1, when cutting point is taken as 7.50, sensitivity was found to be 93% and 1-specifcty was found to be 20%. The reason for determining this value as cutting point in validity analysis was that it was the score in which both sensitive and septicity simultaneously took the best values for Turkish test. In conclusion, it was concluded that the test discriminated the students at a good level and that it was a reliable and valid test.

4. Results and Discussion

Item difficulty indexes showed that difficulty levels of items were different. Therefore, KR-20 reliability analysis was used to examine internal consistency of the test and it as concluded that the Turkish test was reliable. The items showed a moderate level of discrimination. It was concluded that the test was a usable test. An analysis of the cutting point, sensitivity and specificity values determined for the Turkish test using ROC analysis revealed that the area below ROC curve in validity analysis could classify successful/unsuccessful students at a good level. We concluded that ROC analysis can be used as an alternative to other validity analyses. This result is consistent with the findings of Aydemir et al., (2006); Başgül, Etiler, Coşkun, Karakaya and Ağaoğlu (2009). Future studies can be made on development and application of measurement tools in education using ROC and the calculations below ROC curve, which is used in the literature to obtain validity proof in inventory development and adaptation.

References

Aiken, LR. (1994). Psychological testing and assessment. Boston: Allyn and Bacon.

Aydemir, Ö., Demet, M. M., Danacı, A. E., Deveci, A., Taşkın, E. O., Mızrak, S.,et al. (2006). Borderline kişilik envanterinin Türkçe'ye uyarlanması, güvenilirlik ve geçerliliği. *Türkiye'de Psikiyatri Dergisi*, 8 (1), 6-10.

Başgül, Ş. S., Etiler, N., Coşkun, A., Karakaya, I., & Ağaoğlu, B. (2009). Erken çocukluk envanteri- 4 ebeveyn formu Türkçe uyarlamasının geçerlik ve güvenirlik çalışması. *Çocuk ve Gençlik Ruh Sağlığı Dergisi*, 16 (2), 83-93.

Carter, V., & Good, E. (1973). Dictionary of education. New York: McGraw Hill Book Company.

EARGED. (2003). *PIRLS 2001 Ulusal Nihai Rapor.* Ankara: MEB.

Grove,W. M. (2006). *Mathematıcal aspects of diagnosis*. United States of America, 50-75.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29- 36.

Mullis, I. V., Martin, M. O., Kenndey, A. M., Trong, K. L., & Sainsbury, M. (2011). *PIRLS 2011 assesment framework.* Boston: TIMSS&PIRLS International Study Center .

Obuchowski,N.A., Lieber, M. L., & Wians, F. H. (2004). ROC curves in clinical chemistry: uses, misuses and possible solutions. Clinical Chemistry, 50 (7), 1118-25.

Özçelik, D. A. (1998), *Ölçme ve değerlendirme*, Ankara: ÖSYM Yayınları

Wolman, B. B. (1973). Dictionary of education, New York: Reinhold Company.

Yıldırım, C. (1999). *Eğitimde ölçme ve değerlendirme*. Ankara: ÖSYM.