Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 15 (2011) 592–599

Procedia Social and Behavioral Sciences

WCES-2011

CIPP evaluation model scale: development, reliability and validity

Karatas Hakan [a] *, Fer Seval[b]

[a]Yildiz Technical University Besiktas, 34349, Istanbul, Turkey
[b]Yildiz Technical University Besiktas,34349, Istanbul, Turkey

Abstract

The purpose of this study was to determine the validity and reliability of the evaluation scale developed by the researcher based on the principles of Stufflebeam's CIPP Evaluation Model (1988) within the context of the evaluation of English curriculum of Yildiz Technical University. While the scale preparation, by taking advantage of the theoretical information and principles of CIPP Evaluation Model the rough scale composing 65 items was generated and the opinions of the experts were asked for the content and face validity of the scale. Finally, the scale was administered to a large group of students (n=415) from different faculties of Yildiz Technical University in Istanbul. Like the original form, the results of factor analysis indicated that the CIPP Evaluation Model Scale had four factors consisting of context, input, process and product and consisted of 46 items. It was evident from the results that the scale called CIPP Evaluation Scale (CIPP) as a valid and reliable curriculum evaluation instrument can be used in the field of education.

Keywords: CIPP model, context evaluation, input evaluation, process evaluation, product evaluation, curriculum;

1. Introduction

From the elementary education to the level of high education, taking the money, effort and cost spent on foreign language instruction into account it is seen that the results achieved are not satisfactory. The inadequacy of the facilities required for foreign language acquisition and the existence of negativity influencing the language acquisition process might be among the reasons of this condition. Not having been able to analyze this negativity has brought about inaccurate solution suggestions put forth. Actually, taking advantage of contemporary conditions and facilities efficiently can contribute to the solution of this problem (Aslan, 2003). It is known that the aim of foreign language teaching at universities is to be able to follow up scientific resources and write scientific articles. However, it has been observed that even after university graduation, students are still in search of learning foreign languages despite taking English classes during university education. This ascertainment brings forward the question of whether the implemented curriculums are satisfactory or not. Besides, this observation introduces the necessity of regular evaluation process to judge the competence of implemented curriculum. Evaluation in education is realized to define, clarify and conduct a criteria and based on the criteria, to find out objective value, quality, benefit, performance and importance of the evaluation (Worthern, Sanders & Fitzpatrick, 1997; Middlewood & Burton, 2001). Moreover, evaluation is a continuous process which is applicable in planned or unplanned conditions (Gilchrist & Roberts, 1974; Kelly, 1999; Hamilton, 1976) and it aims to question the value of an object

* Karatas Hakan.

1877–0428 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.sbspro.2011.03.146

systematically (Sanders, 1994). To sum up, evaluation is a sophisticated concept which includes the phases of selecting the information, obtaining, analyzing, transferring, using and making a decision on the quality of the curriculum. This study aims to develop a scale to evaluate the English curriculum at university level. Evaluation serves to diagnose the strengths and weaknesses of the curriculum before the implementation and determine the efficiency of the result after the implementation. The aim of collecting data related with the strengths and weaknesses of the curriculum is to ensure the programmers to decide whether the curriculum should be revised, compared, continued or completed (Ornstein & Hunkins, 1988). Due to the diversity in curriculum development, it is not easy to suggest a single model for curriculum evaluation. In the search for curriculum evaluation, researchers can choose the most appropriate model for their aims and conditions or develop a model benefiting from these models (Erden, 1995). It is crucial to be aware of the fact that the adopted evaluation model and accepted program planning model must comply with each other. Otherwise, there will be incongruity and distortion between these models (Kelly, 1999).

Worthern, Sanders and Fitzpatrick (1997) classify evaluation approaches into six groups such as objectives-oriented, management-oriented, consumer-oriented, expertise-oriented, adversary-oriented and participant-oriented approaches. Management-oriented evaluation approach is one of the most important approaches serving managers who are responsible for planning, implementing and evaluating programs. In education, management-oriented evaluation approach provides managers with the information about the implemented program. Hence, the information obtained from evaluation must be the essential part of the decision process and evaluators must contribute to education serving managers, school administrations, teachers and people who need evaluation in education. In this approach, the objectives of the program are not the focal point of evaluation. Stufflebeam has been the pioneer of management-oriented evaluation approach so as to help managers be able to make correct decisions about the program (Worthern, Sanders & Fitzpatrick 1997). His evaluation approach is known as Context, Input, Process and Product Evaluation Model (CIPP). Since 1965, the CIPP evaluation model has been extensively developed and widely implemented (Candoli, Cullen, & Stufflebeam, 1997; Gally, 1984; Granger, Grierson, Quirino, & Romano, 1965; Guba & Stufflebeam, 1968; Nevo, 1974; Stufflebeam, 1969, 1995, 1997-a, 2003-b; Stufflebeam, Candoli, & Nicholls, 1995; Stufflebeam, Gullickson, & Wingate, 2002; Stufflebeam & Millman, 1995; Stufflebeam & Nevo, 1976; Stufflebeam, & Webster, 1988; Webster, 1975).

The aim of the CIPP model attaching importance to process evaluation is to look into all the strategies and components of evaluation and to seek the answers to these questions. Is evaluation design functioning properly? Which points are possibly the problematic ones and how can they be solved? Are there more efficient ways to collect the data (Gilchrist & Roberts, 1974)? Stufflebeam suggests evaluators to follow these steps, as a logical structure, to be used in designing each evaluation type: focusing the evaluation, collection of information, organization of information, analysis of information, reporting of information and administration of evaluation (Wiles & Bondi, 2002; Stufflebeam, 1973, e.g., Worthern, Sanders & Fitzpatrick 1997).

One of the strengths of CIPP model is, especially, that it is a useful and simple tool for helping evaluators produce questions of vital importance to be asked in an evaluation process. Evaluators can determine lots of questions for each component of the CIPP model. Harrison (1993) emphasizes that the CIPP model enables evaluators to intervene the evaluation process when needed, both before and during the program and it also gives the possibility of evaluation for only one component. The CIPP evaluation model has some weaknesses, too. A potential weakness of this model is the evaluator's occasional inability to respond to some significant questions or issues. In planning evaluation procedures, evaluators need to consider the resources and time available. If this model requires more time or resources than are available, another model may have to be considered. (Worthern, Sanders & Fitzpatrick 1997).

2. Method

In this research, the opinions of students regarding the English curriculum were analyzed. Due to reflecting what really exists in the evaluation of four components of the curriculum in this research, as a research model the descriptive research model, which Karasar (1999) expresses as an explanation of an existing situation to give just a

picture and Hovardaoğlu (2000) states as an explanation what the features of some events, beings and groups of people are singly explained or how two or more of their attributes function on a relational level, was implemented.

2.1. Universe and Sample Group

The universe of this research was composed of students of English curriculum. While choosing students' sampling group, the method of cluster random sampling was used. In this research, faculties were determined as clusters and the classes from all the faculties were chosen according to the rate of their representation of the universe composed of the students from different disciplines. The opinions of 415 students in these classes were used in this research.

2.2. Data Collecting Instrument

In this section, the preparation of CIPP scale, the validity and reliability processes of CIPP scale take place. In order to find out answers to the questions of the research, the scale was developed by the researcher in the frame of Stufflebeam's CIPP evaluation model principles. The questions in the scale were in the form of five-point likert scale: (1) I definitely disagree, (2) I disagree, (3) I partly agree, (4) I agree, (5) I completely agree. In the scale, there were 62 items concerning students' opinions about English curriculum. The scale was prepared to cover four components of CIPP evaluation model.

2.3. Analysis of Data

Because the measure instruments were designed as a five-point likert scale, the number value ranging from 1 to 5 was determined for each answer to be able to carry out the analysis. Therefore, the number values used for the options were determined as 5 for "I completely agree", 4 for "I agree", 3 for "I partly agree", 2 for "I disagree" and 1 for "I definitely disagree". The data obtained via the scale was transferred into the computer and the calculations were made using SPSS (Statistical Package for Social Sciences) 13 program as statistical techniques. For data analysis, the means, frequency and standard deviation of the opinions of the students were found. The statistics obtained were transferred into the tables by grouping and then interpreted.

3. Findings

3.1. Validity of CIPP Scale

The initial version of the scale obtained was then sent to five expert English teachers. These teachers were chosen because they were actively involved, for instance, in teaching English in prep classes, in designing and presenting new educational materials in English teaching, or in a testing office of English prep classes. These five teachers were asked to comment on the general structure of the set of statements, and to make comments and suggestions about specific items. In particular, they were asked to consider whether each statement could be scored by students. Their responses varied in length and detail, but, in general, were of a positive and supportive nature. All five made many comments concerning specific statements. These comments were used to make changes in the formulation of almost every statement.

In this research, then it was decided to implement factor analysis for the scale. Before implementing factor analysis, Kaiser-Meyer-Olkin (KMO) test was administered to determine the construct validity and measure the sampling adequacy. The result of KMO test administered for the scale determined the value of P as 0.94. Bayram (2004) specifies that the closer the value of P in KMO test to 1.00 is, the more convenient it is to apply factor analysis for the sampling group. If the value of P in KMO test is lower than 0.50, it is not convenient to administer factor analysis. Barlett's Test of Sphericity is a preliminary test conducted to determine if three or more independent

samples are homogenous or variant before proceeding with an analysis of variance. Barlett's test under taken for sphericity of data showed the value of P as 0.00 (14482.34, sd: 1891, p: 0.00). Since the value of P in KMO test was greater than 0.05 and value of P for Barlett's test was smaller than 0.05, validity of test, sampling adequacy and the factor analysis administered was confirmed.

Factor analysis was started with sixty-two items in the original form of the scale. It was found that the original form of the scale collected in twelve factors whose eigenvalues were greater than one, yet these factors couldn't be given any meaningful names. While preparing the items for the scale, the items were considered in four dimensions, that is, four factors. These dimensions consisted of context, input, process and product evaluation. Therefore, the scale was analyzed in four factors and an exploratory factor analysis with varimax rotation was conducted to examine construct validity of the scale (Figure 1).

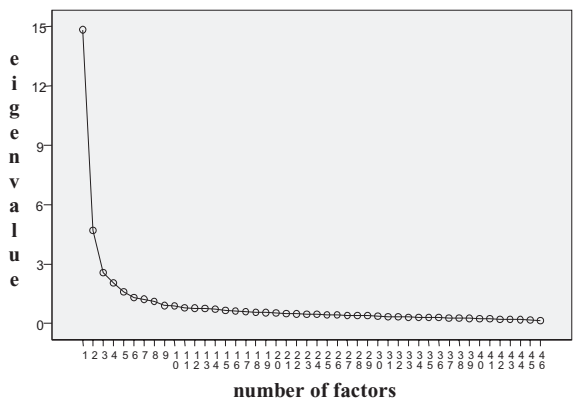

Figure 1. Factor Line Graph

As Büyüköztürk (2002) emphasizes, in a condition in which the same item has high factor loading in two different factors, the difference must be at least 0.10, the factor loadings of the items must be at least 0.45 and the items must be grouped under a single factor. All these things considered, the sixteen items (3, 12, 14, 15, 16, 18, 19, 23, 32, 33, 34, 35, 36, 44, 51, 57) whose factor loadings were under 0.45 and collected under two different factors with high factor loadings were removed from the scale. The results of factor analysis for the scale addressed the four dimensional constructs with 46 items. After removing the items mentioned above from the scale, KMO test was readministered and the results of KMO test determined the value of P as 0.94. Hence, applying factor analysis for the scale was confirmed. Validity analysis is reported in Table 1.

Table 1. Factor Loadings of CIPP Scale After Varimax Rotation

Factor 1 Eigenvalue= 8.98 Variance%= 19.54			**Factor 2 Eigenvalue= 5.56 Variance%= 12.08**			**Factor 3 Eigenvalue= 4.88 Variance%= 10.62**			**Factor 4 Eigenvalue= 4.68 Variance%= 10.19**		
IN	CV	FL	IN	CV	FL	IN	CV	FL	IN	CV	FL
61	.35	.77	1	.45	.71	42	.36	.82	27	.29	.73
62	.38	.75	20	.51	.65	46	.32	.80	26	.26	.72
55	.38	.74	24	.55	.63	43	.37	.79	30	.32	.72
59	.24	.74	22	.53	.57	40	.35	.79	25	.27	.72
52	.36	.70	10	.52	.54	47	.39	.77	29	.37	.70
58	.27	.69	9	.48	.52	41	.41	.76	28	.33	.69
60	.43	.68	17	.67	.49	38	.41	.74			
54	.49	.67	13	.68	.49	39	.41	.73			
56	.53	.65	21	.45	.49						
50	.35	.61	11	.53	.49						

53	.46	.57
7	.40	.59
45	.48	.55
6	.40	.55
31	.48	.52
49	.55	.52
4	.41	.51
8	.52	.51
48	.50	.50
37	.42	.48
2	.44	.46
5	.47	.45

IN: Item number CV: Covariance YD: Factor loading

As it is seen in Table 1, as a result of varimax rotation, the covariance of the items ranged from 0.24 to 0.68 and their factor loadings ranged from 0.46 to 0.82. The total variance explained by four factors was %52.44 (%19.54 by the first factor, %12.08 by the second factor, %10.62 by the third factor and %10.19 by the fourth factor). As the result of varimax rotation, the factor groupings and the dimensions these factors intended to measure are shown in Table 2.

Table 2. Distribution of the Factors According to Their Names and Items

Factors	Competences	Items
1. Factor	Product Evaluation	2, 4, 5, 6, 7, 8, 31, 37, 45, 48, 49, 50, 52, 53, 54, 55, 56, 58, 59, 60, 61, 62
2. Factor	Process Evaluation	38, 39, 40, 41, 42, 43, 46, 47
3. Factor	Input Evaluation	25, 26, 27, 28, 29, 30
4. Factor	Context Evaluation	1, 9, 10, 11, 13, 17, 20, 21, 22, 24

As it is viewed in Table 2, the total number of items belonging to four factors is 46, as 22 items in the first factor, 8 items in the second factor, in the third factor and 10 items in the fourth factor. Analyzing the items of the factors in the scale, the factors were given meaningful names considering the common features of the items grouped under the factors and competences aimed to be measured. 22 items under the first factor were named as “Product Evaluation”, 8 items under the second factor as “Process Evaluation”, 6 items under the third factor as Input Evaluation and 10 items under the fourth factor as “Context Evaluation”.

3.2. Reliability of CIPP Scale

Examining whether each factor was measuring a single idea and whether the items that made up the factors were internally consistent, internal reliability data was obtained through Cronbach’s alpha coefficient among the 4 factors. This is a measure of internal consistency and based on the hypothesis that the measuring instrument is composed of the independent units to fulfill a definite purpose and these units have equal weights that are already known (Karasar, 1999). The scale consists of independent components of the program such as context, input, process and product. In Table 3, the Cronbach’s alpha reliability coefficients and the values of item-total correlation of the components are shown.

Table 3. Cronbach’s Alpha Reliability Coefficients and Item-total Correlations of the Components

Competences	Item Numbers	A	r
Context Evaluation	1, 9, 10, 11, 13, 17, 20, 21, 22, 24	.81	.42 - .59
Input Evaluation	25, 26, 27, 28, 29, 30	.87	.64 - .70
Process Evaluation	38, 39, 40, 41, 42, 43, 46, 47	.92	.69 - .77
Product Evaluation	2, 4, 5, 6, 7, 8, 31, 37, 45, 48, 49, 50, 52, 53, 54, 55, 56, 58, 59, 60, 61, 62	.94	.51 - .77

A: Alpha values r: Item-total correlation

As it is shown in Table 3, the reliability coefficients of four components ranged from 0.81 to 0.94 and the reliability coefficient of the whole scale was found as 0.95. A reliability coefficient of .70 or higher is considered "acceptable" in Social Science research situations. The coherence level increases when reliability coefficient approaches to 1.00, and decreases as it approaches to 0 (Turgut, 1997; Yıldırım, 1999). In accordance with the reference mentioned above, the internal consistency indicated reasonably high reliability. Internal reliability on each scale was also obtained via item-scale correlation. Findings showed that the four factors had internal consistency reliabilities ranging from 0.42 to 0.77. Pearson's correlations that were significant at the 0.01 level indicated acceptable internal reliability.

Internal consistency reliability of scale items was evaluated with item-total correlation and total correlation coefficients between points that individuals got for each item of scale, average deviation, standard deviation and item-total correlation findings have been given in Table 4.

Table 4. Values of mean, standard deviation and corrected item-total correlations of the scale

IN	X	sd	r	IN	X	sd	r
1	3.86	.97	.59	37	3.09	.93	.65
2	3.33	.94	.58	38	4.13	.94	.69
4	3.05	1.02	.63	39	3.79	1.01	.70
5	2.88	1.04	.56	40	3.70	1.05	.74
6	3.00	.92	.65	41	3.98	.92	.73
7	3.18	1.04	.65	42	4.13	.99	.60
8	2.77	1.09	.58	43	3.87	1.02	.57
9	3.14	.89	.48	45	3.00	.95	.62
10	3.28	.95	.42	46	3.94	.98	.63
11	3.09	.87	.45	47	3.93	.95	.55
13	3.53	1.05	.42	48	3.11	1.07	.57
17	3.28	.95	.48	49	2.70	1.13	.51
20	3.70	.89	.57	50	3.15	.97	.72
21	2.76	1.00	.50	52	3.10	1.01	.73
22	3.32	.88	.57	53	3.22	1.01	.60
24	3.39	.94	.58	54	2.77	1.08	.60
25	2.87	1.21	.69	55	2.65	1.07	.68
26	2.66	1.12	.70	56	2.94	1.02	.61
27	2.83	1.09	.70	58	3.11	.97	.74
28	3.13	1.09	.64	59	3.06	.97	.77
29	2.72	1.06	.66	60	2.44	1.07	.65
30	3.13	1.09	.66	61	2.47	1.07	.65
31	3.16	.91	.63	62	2.53	1.08	.67

IN: Item Numbers X: Mean SS: Standard Deviation r: Corrected item-total correlation

As it can be seen in Table 4, internal consistency reliability of scale articles was also evaluated with item-total correlation and total correlation coefficients between points that individuals got from each item of survey and dimensions were given values between 0.42 and 0.77. In all items, positive and meaningful correlation (p=01) were found. On any scale, items with values more than 0.20 are evaluated as acceptable and values more than 0.30 are evaluated as good level. Moreover, although there isn't any exact rule, item-total correlations mustn't be negative for additivity feature of scale. (Turgut,1997). In line with references mentioned above, item-total correlations can be considered as good level.

4. Results

In this research, Stufflebeam's CIPP evaluation model has been implemented. The reason why this model has been preferred is that it is feasible in foreign languages curricula and involves various evaluation types such as context, input, process and product evaluation. With context evaluation component, it is aimed to determine capabilities of subjects such as program's convenience to development of students' linguistic skills, balancing their linguistic skills, having measurable criteria, adequacy of program time and convenience to difficulty levels of subjects, school books' convenience to students' levels, catching attention of students and understandability of

subjects. With input evaluation component, it is aimed to find answers to questions such as to what extent visual and audial materials and written materials (study papers) used in the program make it easier for students to learn, whether they catch attention of them and whether they have positive effects on their language skills. With process evaluation component, it is aimed to find answers to questions such as having sufficient exercises and revisions, having homework with qualifications sufficient enough to support learning, students' attending the lesson actively, having enough exams, having exercises suitable for double and triple group studies and using all language skills and teachers' spending enough time to solve students' problems during program implementation step. With product evaluation component, it is aimed to find answers to questions such as meeting the individual needs, characteristics, interests of students completing the program, meeting the available and future needs of students, contribution to students in works related with their departments, encouraging students to learn English, having projects with positive effect on students' language skills, getting them adopt the habit to study, giving students the opportunity to use what they learned, increasing vocabulary knowledge, getting them adopt the habit to study with a group, having satisfactory outputs of reading, writing, speaking and listening skills and grammar levels, students' getting satisfactory information and skills related with language, and getting information related with English required to complete the department lessons of students and necessary for their departments and several work areas.

Consequently, the results of the present research may be of interest to researchers, educators, and to the educational process. That is, the scale may be used as an efficient instrument in order to evaluate English curriculum at any stage. This may open new perspectives in the field of evaluation and assessment of English curriculum. The results of this study also point to the direction for future researches.

References

Aslan, O. (2003). Türkiye'de yabancı dil öğretme ve öğrenmede karşılaşılan sorunları çözümlemeye yönelik bir öneri: Yabancı dil öğretimi ve öğreniminde internet kullanımı. Akademik Bilişim Konferansı, Çukurova Üniversitesi.

Bayram, N. (2004). Sosyal bilimlerde SPSS ile veri analizi. Bursa: Ezgi Kitabevi.

Candoli, I. C., Cullen, K., & Stufflebeam, D. L. (1997). Superintendent performance evaluation: Current practice and directions for improvement. Boston: Kluwer.

Gally, J. (1984, April). The evaluation component. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Gilchrist, R. S. and Bernice R. R. (1974). Curriculum development (A humanized Systems approach). California: Lear Siegler, Inc./Fearon Publishers.

Granger, A., Grierson, J., Quirino, T. R., & Romano (1965). Training in planning, monitoring, and evaluation for agricultural research management: Manual 4–Evaluation. The Hague: International Service for National Agricultural Research.

Guba, E. G., & Stufflebeam, D. L. (1968). Evaluation: The process of stimulating, aiding, and abetting insightful action. In R. Ingle & W. Gephart (Eds.), Problems in the training of educational researchers. Bloomington, IN: Phi Delta Kappa.

Erden, M. (1995). Eğitimde program değerlendirme. Ankara: Pegem Yayıncılık.

Hamilton, D. (1976). Curriculum evaluation. London: Open Books Publishing Limited.

Harrison, A. S. (1993). An evaluation model for middle school counseling and guidance. Old Dominion University.

Hovardaoğlu, S. (2000). Davranış bilimleri için araştırma teknikleri. Ankara: VE-GA Yayınları.

Karasar, N. (1999). Bilimsel araştırma yöntemi. Ankara: Nobel Yayın Dağıtım.

Kelly, A. V. (1999). The curriculum. London: Paul Chapman Publishing Limited.

Middlewood, David and Neil Burton (2001). Managing the Curriculum. London: Paul Chapman Publishing.

Nevo, D. (1974). Evaluation priorities of students, teachers, and principals. Unpublished doctoral dissertation, The Ohio State University, Columbus.

Ornstein, A. C. and Hunkins F. P. (1988). "Curriculum: Foundations, principles, and issues". New Jersey: Prentice Hall.

Sanders, J. R. (1994). The program evaluation standarts. California: SAGEPublications, Inc.

Stufflebeam, D. L. (1969). Evaluation as enlightenment for decision-making. In A. Walcott (Ed.), Improving educational assessment and an inventory of measures of affective Behavior. Washington, DC: Association for Supervision and Curriculum Development.

Stufflebeam, D. L., & Nevo, D. (1976, Winter). The availability and importance of evaluation information within the school. Studies in Educational Evaluation, 2, 203-9.

Stufflebeam, D. L., & Webster, W. J. (1988). Evaluation as an administrative function. In N. Boyan (Ed.), Handbook of research on educational administration (pp. 569-601). White Plains, NY: Longman.

Stufflebeam, D. L. (1995). Evaluation of superintendent performance: Toward a general model. In A. McConney (Eds), Toward a unified model of educational personnel evaluation. Kalamazoo: Western Michigan University Evaluation Center.

Stufflebeam, D. L., & Millman, J. (1995, December). A proposed model for superintendent Evaluation. Journal of Personnel Evaluation in Education, 9(4), 383-410.

Stufflebeam, D. L., Candoli, C., & Nicholls, C. (1995). A portfolio for evaluation of school superintendents. Kalamazoo: Center for Research on Educational Accountability and Teacher Evaluation, The Evaluation Center, Western Michigan University.

Stufflebeam, D. L. (1997-a). Strategies for institutionalizing evaluation: revisited. Occasional Paper Series #18. Kalamazoo: Western Michigan University Evaluation Center.

Stufflebeam, D. L. (2003-b). Institutionalizing evaluation in schools. In T. Kellaghan & D. L. Stufflebeam (Eds.), The international handbook of educational evaluation (Chapter 34). Boston: Kluwer.

Stufflebeam, D. L., Gullickson, A. R., & Wingate, L. A. (2002). The spirit of Consuelo: An evaluation of Ke Aka Ho‘ona. Kalamazoo: Western Michigan University Evaluation Center.

Turgut, M.F. (1997). Eğitimde ölçme ve değerlendirme metodları. Ankara: Gül Yayınevi.

Webster, W. J. (1975, March). “The organization and functions of research and evaluation in large urban school districts.” Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Wiles, J. and Bondi J. (2002). Curriculum development-A guide to practice-. New Jersey: Merrill Prentice Hall.

Worthern, B. R., Sanders J. R. and Fitzpatrick J. L. (1997). Program evaluation – Alternative approaches and practical guidelines-. New York: Longman, Inc.

Yıldırım, C. (1999). Eğitimde ölçme ve değerlendirme. Ankara: ÖSYM yayınları, 1999-4.