Education and Science

Original Article

Vol 49 (2024) No 217 225-248

Development of a Meta-Evaluation Rubric and Meta-Evaluation of Initial Teacher Education Programs

Sevinç Gelmez Burakgazi [1], Yasemin Karsantık [2]

Abstract

In this study, evaluation research in initial teacher education programs were evaluated with a rubric developed in line with Joint Committee Standards (Yarbrough, Shulha, Hopson, & Caruthers, 2011) This meta-evaluation study consisted of two phases. In the first phase, a rubric was developed to assess the evaluation reports based on program evaluation standards. In the second phase, theses and articles selected with certain criteria were analyzed based on the meta-evaluation rubric. Adopting the empirical reevaluation of multiple data sets about the same program model, the data were analyzed with the descriptive analysis method. According to the results, the selected research mostly met accuracy standard whilst feasibility and propriety standards were limitedly met. It was concluded that program evaluation research in the Turkish context needed to be improved by further considering program evaluation standards.

Keywords

Joint committee standards
Program evaluation
Program evaluation standards
Meta-evaluation
Initial teacher education programs

Article Info

Received: 03.27.2022
Accepted: 10.09.2023
Published Online: 12.15.2023

DOI: 10.15390/EB.2023.11755

Introduction

Evaluation conveys distinctive meanings which may be outcome-oriented like determining whether the objectives are achieved, feedback-oriented like providing evidence for decision-making, or quality-oriented like assessing the merit and worth (Stufflebeam, 2011). Similarly, program evaluation is a research-based process which aim to identify the level of achieving objectives, to make judgment on the efficacy of curriculum (Oliva & Gordon, 2013), to identify strengths and weaknesses of curriculum (Ornstein & Hunkins, 2017) and to provide data for improving program and practices in education, human services- in virtually every area of society (Fitzpatrick, Sanders, & Worthen, 2004). Curriculum evaluation enables evaluators to judge the merit which is independent of context such as integrity and clarity of curriculum, and worth which is framed by context such as appropriateness to learners (Fitzpatrick et al., 2004; Lincoln & Guba, 1980; Melrose, 1998; Stufflebeam, 2002).

Meta-evaluation as a separate professional field, on the other hand, is evaluation of the program evaluation to meet the needs of quality control with the increase in program evaluation models and studies (Sağlam & Yüksel, 2007). Sound evaluations are complemented with evaluation of efforts for those evaluations; therefore, meta-evaluation is utilized for both improving ongoing evaluations, and determining merit and worth of the completed evaluations (Stufflebeam, 2011). Evaluations are

[1] Hacettepe University, Faculty of Education, Department of Educational Sciences, Türkiye; University College London, Department of Curriculum, Pedagogy, and Assessment, United Kingdom, sevincgelmez@gmail.com
[2] Trabzon University, Fatih Faculty of Education, Department of Educational Sciences, Türkiye, ybaykin@gmail.com

described and judged against the criteria in functional, technically-adequate and cost-effective meta-evaluations (Stufflebeam, 2011). Meta-evaluations may differentiate in terms of the utilized criteria or methods. To explain, standards such as Program Evaluation Standards by Joint Committee on Standards for Educational Evaluation (JCSEE) provide variety in criteria whereas methodological differences may be exemplified as providing strengths and weaknesses of evaluation (Cooksy & Caracelli, 2009).

Conceptual framework

Meta-evaluation: definitions and purposes

The concept 'meta-evaluation' was first used by Michael Scriven as evaluating the evaluations of educational products in 1969; then has turned into a method applied for evaluation of evaluation in the field of educational programs (Stufflebeam, 2000a). Making a distinction between the theoretical and practical functions of meta-evaluation, Scriven (1969, p. 36) defines meta-evaluation as "…the methodological assessment of the role of evaluation" in theoretical terms, and as "… is concerned with the evaluation of specific evaluative performances" in practical terms. To Stufflebeam (2000a, p. 95), meta-evaluation "… is the process of delineating, obtaining, and applying descriptive information and judgmental information -about the utility, feasibility, propriety, and accuracy of an evaluation in order to guide the evaluation and to publicly report its strengths and weaknesses". Stufflebeam (2001) also added "its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility" to the utility, feasibility, propriety, and accuracy dimensions as stated in the aforementioned definition of meta-evaluation and make the definition more comprehensive (p. 205).

Meta-evaluation incorporates formative and summative types with different functions (Stufflebeam, 2000a, 2001, 2011). Whilst formative meta-evaluation deals with improving the quality of evaluation and provides feedback to evaluators to make sound decisions in issues such as the purpose of evaluation, data collection and analysis procedures, summative meta-evaluation provides evidence on the merit of evaluation and whether evaluations meet the quality of standards (Stufflebeam, 1978). In other words, formative meta-evaluation is required to plan and conduct sound evaluations, and summative meta-evaluation assesses the quality of completed evaluations (Stufflebeam, 2001). Formative meta-evaluation may be conducted based on the criteria such as stakeholder participation, sample of meta-evaluation, meta-evaluation management plan and ethical issues. Summative meta-evaluation may be based on the Joint Committee's Program Evaluation Standards (JCPES) (Stufflebeam, 2000a).

Meta-evaluation is of significance in enabling evaluators to ensure quality of their evaluations and providing feedback to improve evaluations (Stufflebeam, 2000a). JCPES contribute to meta-evaluations in terms of offering explicit criteria, thereby being preferred in meta-evaluation studies some of which have been conducted in international context (e.g. Akıncı & Köse, 2020, 2022; Cooksy & Caracelli, 2009; Tingle, DeSimone, & Covington, 2003; Widmer, 2000; Yağan, 2019; Yasar, Gultekin, Kose, Girmen, & Anagun, 2005). Widmer (2000) examined 15 Swiss evaluation studies in the fields such as environment, industry and social policies based on standards set by the JCSEE and concluded that those evaluation studies strongly met the standards (and indicators) of utility (evaluator credibility, information scope and selection, report dissemination, report timeliness), feasibility (practical procedures) and propriety (formal obligation, public's right to know, rights of human subjects); however, they were weak in accuracy standard regarding the indicators of valid measurement, reliable measurement, described purposes and procedures and justified conclusions. Scott-Little, Hamann, and Jurs (2002) conducted a meta-evaluation on after-school programs based on the program evaluation standards set by JCSEE and found that after-school evaluation reports had moderate compliance with program evaluation standards, however, they were limited in terms of research designs.

Tingle et al. (2003) assessed 11 evaluation studies on school-based smoking prevention programs through accuracy standard, which is one of the four guiding principles (feasibility, propriety, accuracy, utility) established by the JCSEE, via three-point rating system in their meta-evaluation. The results demonstrated that evaluations met the criteria about research design and statistical analyses at the highest level whereas they did not meet the criteria about validity and reliability.

Models of meta-evaluation

Cook and Gruder (1978) posit three research traditions in meta-evaluation which are (1) reanalyzing and evaluating different completed evaluations, (2) rating different evaluations in terms of technical competency and (3) providing information for the issues such as proposing the quickest solutions to the unexpected problems. Figure 1 displays models of meta-evaluation developed based on those traditions:

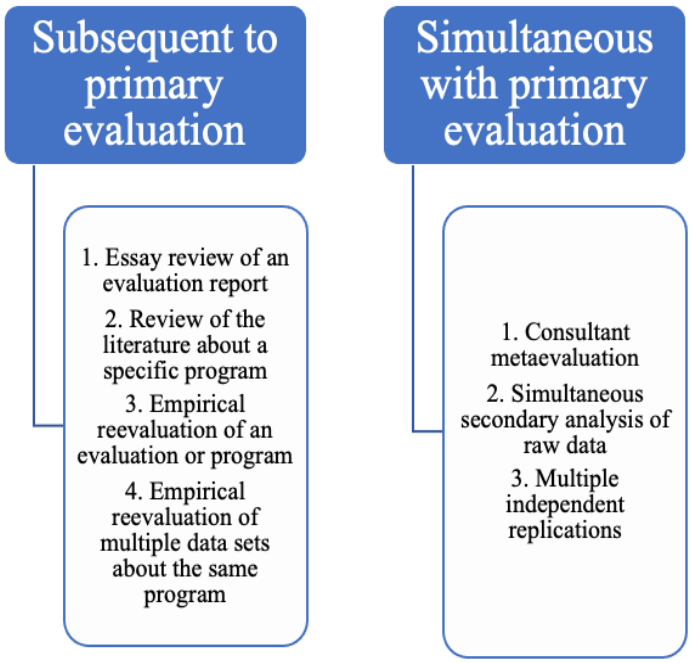

Figure 1. Models of meta-evaluation (Adapted from Cook & Gruder, 1978, p. 17)

Models for subsequent to primary evaluation displayed in Figure 1 are explained by Cook and Gruder (1978, p. 18) as follows:

1. Essay review of an evaluation report points out comments on a single evaluation dataset.
2. Review of the literature about a specific program indicates comments on more than one dataset about a specific program.
3. Empirical reevaluation of an evaluation or program refers to manipulation of a single dataset about a program to determine the validity of primary evaluation results.
4. Empirical reevaluation of multiple data sets about the same program focuses on manipulation of multiple datasets about a program to determine the validity of primary evaluation results.

Models for simultaneous with primary evaluation demonstrated in Figure 1 are described by Cook and Gruder (1978, p. 18) as follows:

1. Consultant meta-evaluation aims to provide feedback to improve an ongoing evaluation and enables monitoring of multiple evaluations about the same program.
2. Simultaneous secondary analysis of raw data means analysis of an ongoing evaluation by individuals outside primary evaluators.
3. Multiple independent replications enable more than one evaluator to design and implement an evaluation independently.

The current meta-evaluation study adopted the empirical reevaluation of multiple data sets about the same program model in that the researchers obtained the research on evaluation of initial teacher education programs and reevaluated those multiple datasets based on JCPES.

Program evaluation standards

Some professional criteria are needed to evaluate the efficiency of services provided in both education and other fields requiring evaluation. To meet the needs, institutions and associations make attempts to develop evaluation standards (Stufflebeam, 2004). Basic issues in meta-evaluation are about standards. It is possible to determine quality of evaluations through making judgments with certain standards (Stufflebeam, 1978). Several standards exist in literature. To exemplify, AEA (American Evaluation Association) Guiding Principles provide criteria for evaluations in a wide range of fields including education (Stufflebeam, 2001). On the other hand, program evaluation standards by JCSEE are possibly more commonly used in educational settings. The Joint Committee's Program Evaluation Standards is the outcome of a project started in 1975. In 1989, the Joint Committee extensively reviewed the process, the standards were combined and new ones were added. What is important here is, with this revision, new illustrations to be able to be used in settings that include schools, universities, law, business, government have been added (Sanders, 1994). The standards, which are developed to be utilized in educational evaluations in the USA and Canada, have achieved the attention of different countries (Stufflebeam, 2004). JCPES, which incorporates 30 standards organized into utility, feasibility, propriety, accuracy and evaluation accountability attributes of evaluation (Yarbrough et al., 2011) focus on evaluations of curriculum and school personnel (Stufflebeam, 2001). The locus of the five standards might be summarized as follows (Yarbrough et al., 2011):

1. Utility standards address whether both evaluation products and processes meet the needs of stakeholders.
2. Feasibility standards focus on increasing the efficiency and effectiveness of the evaluation.
3. Propriety standards promote appropriateness, fairness and legality of the evaluation.
4. Accuracy standards are related to whether the evaluation is accurate and reliable.
5. Accountability standards propose documentation and meta-evaluation of the evaluation.

Yüksel and Sağlam's (2011) study is a unique example of the search for evaluation standards for our country. In their study, the researchers gathered data from 158 faculty members from 94 state universities and based on their reviews, 23 evaluation standards and 110 indicators were defined under four standards: Utility, feasibility, propriety, and accuracy.

Considering evaluation from a wider perspective, JCPES have been established with the endeavors of joint groups such as teachers, evaluators, statisticians, supervisors, administrators and policymakers to empower theories and practices of evaluation, and to present a common language and guidelines (Stufflebeam, 2004). Therefore, in the current study, meta-evaluation of the initial teacher education program evaluation research was grounded in Joint Committee Standards.

Whereas plenty of review studies on program evaluation were conducted in Turkish educational context (e.g. Kablan, 2011), most of them were not designed as meta-evaluations. Considering the meta-evaluations, it is evident that they are grounded in JCPES. Yüksel and Akın (2013) conducted a meta-evaluation study to determine the compliance of 2009 Student Achievement Examination reports with JCPES via a checklist developed based on 27 standards items and analyzed data through document analysis. The results indicated that 11 standards were completely met, 6 standards were partially met, and 10 standards were not met, and the report was partially practical in conducting with a plan, partially useful regarding legal and ethical issues, partially feasible in efficiency and not sufficient in interpretation, justified conclusions and recommendations. Yağan (2019) assessed doctoral dissertations on program evaluation completed between 2015-2018 based on the 4 dimensions, 26 standards and 206 sub-standards of meta-evaluation checklist developed by Stufflebeam (2012). The results demonstrated that standards such as evaluator credibility, project management, participant rights and respect and valid information were met; however, standards such as attention to stakeholders, contextual viability, justified conclusions and decisions, evaluation design and analysis and reporting were not sufficiently met.

There also exist meta-evaluations conducted in Turkish context on evaluations of initial teacher education programs. Yasar et al. (2005) assessed 8 program evaluation studies on teacher training programs for elementary education conducted between 1997-2004 based on feasibility, utility accuracy and propriety standards via three-point rating system and concluded that evaluations were sufficient in feasibility, and partly sufficient in utility, accuracy and propriety. Akıncı and Köse (2020) examined 9 initial teacher training program evaluation studies, in which at least one program evaluation approach or model is utilized, via a checklist developed based on JCPES except accountability standard, and concluded that all indicators under accuracy, feasibility, utility and propriety standards were moderately met; however, proof of some standards did not exist and some of the standards were not met at all. The current study differentiates from aforementioned studies in terms of three points. Firstly, the current study addressed accountability standards, which were not included in those studies, since it seems applicable to analyze research reports based on the information on documentation, internal and external evaluation. By reading the research reports, one can understand whether documentation is detailed enough, and internal and external evaluation is reported or not. Second, the current study aimed not only to evaluate ITE program evaluations based on JCPES, but to develop an instrument (a meta-evaluation rubric) that can be used for several purposes and presented clearly with all items/indicators, as well. Thus, researchers can utilize the instrument to evaluate the program evaluation studies and evaluators can draw on the instrument to check whether their program evaluation studies meet the program evaluation standards. They can see all items/descriptors and can find out which points they will focus on both conducting and evaluating program evaluation. However, in the previous meta-evaluations, the instruments mentioned in the studies were not included. Last but not the least, which makes the study one of the unique evaluation studies is that, the current study addressed 24 evaluation studies conducted between 2010 and 2020 indicating that it is more comprehensive than existing meta-evaluations.

The current study

The purpose of the study was twofold; firstly, to develop a rubric to assess the evaluation reports based on program evaluation standards and secondly to determine to what extent do initial teacher education program evaluations comply with JCPES. The current meta-evaluation study assessed 24 initial teacher education program evaluations conducted between 2010 and 2020 in terms of their compliance with propriety, accuracy, feasibility, utility, and accountability standards set by JCSEE. Most importantly, the study aims to identify the strong and weak points in the evaluations of initial teacher education programs, to reveal insights into the current evaluation research by highlighting the points that need to be improved in evaluation studies. As Gözütok (2006) states, continuity in

curriculum development as questioning and evaluating the effectiveness of the program is a starting point for the development of programs. In parallel with this, the results are expected to provide feedback to improve evaluations as Stufflebeam (2000a) suggested. Stufflebeam (2000b) also asserts that meta-evaluations serve a range of stakeholders from policy-makers to practitioners and students. From this perspective, this meta-evaluation study is expected to contribute researchers to conduct more effective program evaluations, help teacher educators to be more informed about the efficiency of the program they implement and support policy-makers to make informed decisions about ITE programs. Last, those results are supposed to contribute to the program evaluation studies in the international context in terms of identifying which aspects may be omitted in evaluations. Accordingly, the research question is as follows:

1. Do initial teacher education (ITE) program evaluation research conducted in Türkiye in the last decade (2010-2020) meet utility, feasibility, accuracy, propriety, and accountability standards?

Method

Meta-evaluation process requires both social (group) process and technical tasks. Whereas social process may be exemplified as the interaction between meta-evaluator and stakeholders in issues such as mutual understanding of the purpose and questions of meta-evaluation and mutual interpretation of meta-evaluation results, technical tasks include several tasks such as data collection, development of data collection instruments (Stufflebeam, 2001). The current meta-evaluation study is inclined to focus on technical tasks rather than group interactions and processes; thereby revealing the need to develop an instrument for assessing the published program evaluation research based on JCPES. Rather than translating Program Evaluations Metaevaluation Checklist developed by Stufflebeam (2012), a new instrument was developed since the written reports of program evaluation research would be assessed and some items in checklist of Stufflebeam (2012) would not be applicable such as “help stakeholders understand the evaluation plan” (1) and “train staff” (4).

Meta-evaluation is a way to evaluate concluded evaluations in which qualitative data analysis is performed. A new evaluation is executed in meta-evaluation in which evaluations are exposed to analysis based on the criteria mostly defined by the Joint Committee (Hedler & Gibram, 2009). The current meta-evaluation adopted an approach which encompassed descriptive qualitative analysis of ITE program evaluation based on the meta-evaluation rubric developed by the researchers. In this context, the study was conducted in two stages: Meta-evaluation rubric development and document review. Those stages were further explained below:

Development of meta-evaluation rubric

For the first part of the study, a meta-evaluation rubric was developed. In order to develop the rubric, the steps suggested by Wolf and Stevens (2007) was followed:

1. Identification of Performance Criteria: "The first step in developing a rubric is to identify the criteria that define the performance" (Wolf & Stevens, 2007, p.5). As we based our research on the Program Evaluation Standards book by the Joint Committee on Standards for Educational Evaluation (3rd edition) (Yarbrough et al., 2011), the standards were examined carefully to determine the rubric items. In this study, the items were gathered under five categories as suggested by the Committee: utility, accuracy, feasibility, propriety and accountability.
2. Setting Performance Levels: "The second step in the process is to decide how many levels of performance are appropriate for the assessment." (Wolf & Stevens, 2007, p.5). There are typically three to six rating levels in the rubrics. The choice of the number is decided considering the purpose for the assessment. A lower number of levels would be more appropriate if the primary purpose was to make a summative decision, e.g. whether someone passed or did not pass their course and took an exam for example. In our case, we used three levels: "Yes", "No" and "Partially" to make a decision. (please see Table 1)
3. Creating Performance Descriptions: "The third step in the process is to write a description for each cell in the matrix." (Wolf & Stevens, 2007, p.7). This points to a brief paragraph providing sufficient information to guide the scoring but without overwhelming the readers / performers (Wolf & Stevens, 2007). We used the "Notes / Evidences (Please elaborate your response)" column to give details to the readers / performers.

The presence of performance criteria and performance level descriptors differentiate rubrics from the other data collection tools. Table 1 displays the meta-evaluation rubric.

Table 1. The meta-evaluation rubric

Information about the study / Type of the study (article, thesis, etc.)					
Criteria		Performance Levels			Performance Descriptors
Standards (Criteria)	Items/questions	Yes	No	Partially	Notes / Evidences (Please elaborate your response)

Moskal and Leydens (2000) suggested examining content, construct and criterion validity to support the validity of the rubric. Content - related evidence was ensured by preparing the rubric in accordance with the Program Evaluation Standards. For construct - related evidence, scoring criteria was determined as "Yes", "No" and "Partially" to be able to evaluate all of the important aspects of the intended construct as discussed by Moskal and Leydens (2000). In addition to that, the key components of relevant performance that can be assessed through the use of the assessment tool are included (e.g. under the feasibility standard, item 4: Have independent (external) expert opinions been taken in the evaluation process?) as a way to serve criterion validity.

The draft rubric was submitted to expert opinion in the context of validity and reliability of the tool and process. Three professors working at the Curriculum and Instruction departments at three different public universities were sent forms and they were asked to examine the draft rubric. The content was found to be appropriate with the standards, but some minor changes were made (e.g., rewriting some items, deleting three items). For instance, expert 1 suggested deleting three indicators in the utility standards in that they were about individual and cultural values and they were unclear whether values were reflected into the aim of evaluation or they were the aims themselves. Another example is that expert 3 suggested combining two draft indicators in the utility standards since they

were about stakeholders' needs. Therefore, they are rewritten as "Has the evaluation been designed to meet the needs of stakeholders?"in the final form of the rubric. Lastly, the rubric was piloted with evaluation research and the form was revised and prepared for the final version.

The final form consisted of 53 items rating a single choice (Yes/No/Partial). *Yes* statement refers to fully meeting the standard whereas *no* statement means not meeting the standard at all. However, *partially* statement refers to there is some proof that the standard is met but it is not adequate. They were exemplified in the results section.

Document review

For the second phase of the study, document review was conducted. Document review includes the selection of documents (superficial review), reading (comprehensive review) and interpretation processes (Bowen, 2009). According to Yıldırım and Şimşek (2016), "Document analysis in qualitative research can be a data collection method alone or can be used together with other data collection methods" (p.187). Bowen (2009) addressed the strengths of document analysis as being effective, easy to access data, economical, unaffected by the research process and the presence of the researcher, and it is a data collection method based on the examination of documents containing comprehensive details including author information and other details. Besides its strengths, document review also has its weaknesses which may be access barriers, bias in the selection of documents and the nature of documents such as not being produced for research purposes (Bowen, 2009). To avoid the weaknesses in document analysis, the researchers took some precautions. Scientific research was obtained in several searching and indexing bases, and the criteria considered in document selection were clearly presented.

Data analysis

The documents obtained within the scope of the study were analyzed with the descriptive analysis method. The descriptive analysis was carried out in four stages: Organization of the documents based on the criteria (database, time interval, subject etc.), categorization of each document (theses and articles) around the framework separately, describing and interpreting the findings with quotations where necessary. "In the end, whether a given standard has been addressed adequately in a particular situation is a matter of judgment. Precise decision rules regarding satisfaction of a standard cannot be specified. Such rules would be arbitrary and not universally applicable; they would likely delude and mislead. Nevertheless, evaluators who cite the standards should describe clearly how they used them." (Sanders, 1994, p.9). In the light of the given rationale, in the results part, we provided evidence to support our judgments under the tables. Considering ethical procedures, abbreviations were used for the thesis and articles used for analysis (e.g. T1, T2, T3…for thesis; A1, A2, A3 for articles)

Data collection procedures

In this study, the keywords "program evaluation curriculum evaluation" in searching and indexing bases such as "Google Academic, ULAKBIM, CoHE (Council of Higher Education) National Thesis Center, Scopus, EBSCOhost, ERIC, JStor, Web of Science, Proquest, Thomson Reuters, Wiley and Science Direct" with search made between 2010-2020 years and in the context of Türkiye 27 studies have been reached. However, three theses were published as an article in national and international journals. Therefore, those theses were omitted, and their article forms were included in the study. Overall, four thesis and 20 articles were analyzed.

Validity and reliability

Validity and reliability are key aspects to ensure the quality of a qualitative research and mostly can be named together as trustworthiness. Trustworthiness is established with four components: credibility, transferability, confirmability and dependability. In the study, the role of the researchers was described, the processes of data collection, data analysis and their interpretations were explained in a consistent and detailed manner without the participation of comments, and the whole process was followed by two researchers in order to increase credibility and transferability. In addition, an expert from the Department of Curriculum and Instruction was asked to analyze four research randomly determined by the researchers according to the determined thematic framework and Miles and Huberman (1994) Reliability = Consensus / (Consensus + Disagreement) x 100 formula, according to which 89% consensus was calculated.

Researchers' roles and experiences in program evaluation

Researchers' expertise and experiences in program evaluation is one of the key factors for meeting utility standards in an evaluation study. For the current meta-evaluation, both authors had an expertise in curriculum development and evaluation with national and international experiences. They graduated from research universities, faculties of education and held a PhD degree in curriculum and instruction. Therefore, they both have knowledge and experiences on the courses in ITE programs and program evaluation. They also instructed courses on program evaluation in both undergraduate and graduate levels. As they both work as teacher educators in faculties of education and curriculum and instruction departments, they are well informed about the courses in ITE programs. Moreover, the first author conducted program evaluation research in one of the courses in ITE programs and the second author carried out descriptive research on the relationship between teachers and program evaluation. Regarding their roles in this meta-evaluation, the authors both read the handbook of JCSEE in detail, wrote relevant items and checked their clarity, and conducted the analysis of articles and theses based on JCPES to ensure coherence and objectivity.

Results

Results of the study were presented in line with the research question following utility, feasibility, accuracy, propriety and accountability, respectively.

Do ITE program evaluation research conducted in Türkiye in the last decade meet utility standards?

To address the utility standards, the evaluation research were analyzed to find out whether they incorporate the indicators of utility standards. Table 2 displays the results.

Table 2. Results on utility standards

Standard	Items/Questions		Yes	No	Partially
Utility	1. Have the evaluator(s)' expertise in program evaluation been reported?		6	14	4
	2. Have the evaluator(s)' experiences in program evaluation been reported?		2	22	0
	3. Have the evaluator(s)' experiences relevant to the evaluated program been reported?		4	20	0
	4. Have the stakeholders who may be affected by the evaluation (such as experts, participants) been identified?		5	1	18
	5. Have the stakeholders who may affect the evaluation (such as experts, participants) been identified in the study?		9	0	15
	6. Have the participants been informed about the purpose of the evaluation?		6	18	0
	7. Have the evaluation data been consistent with the purpose of evaluation?		23	1	0
	8. Have the evaluation data been collected from different sources (such as teachers, parents, school principals)?		13	10	1
	9. Has the evaluation been designed to meet the needs of stakeholders?		3	19	2
	10. Has the purpose of the evaluation been reported explicitly and clearly?		24	0	0
	11. Has the evaluation process been reported explicitly and clearly?		7	13	4
	12. Have the evaluation results been reported explicitly and clearly?		24	0	0
	13. Have the evaluation results been reported in a meaningful way for the stakeholders (in a non-technical language)?		20	4	0
	14. Has the communication between evaluator(s) and stakeholders been reported in the evaluation process?		3	20	1
	15. Have the potential benefits or effects of evaluation results for stakeholders been reported?		6	14	4
	16. Has the evaluation been justified in the relevant conditions and context?		19	4	1
	Total	**f**	**174**	**160**	**50**
		%	**46%**	**42%**	**12%**

As indicated in Table 2, the ITE program evaluation research were found to meet the utility standard at the rate of 46 percent and partially meet the relevant standard at the rate of 12 percent. However, the program evaluation research did not meet the utility standard at the rate of 42 percent. Whereas two descriptors (*Has the purpose of the evaluation been reported explicitly and clearly?* and *Have the evaluation results been reported explicitly and clearly?*) were completely met in all the evaluation research included in the meta-evaluation (n=24), it was observed that the descriptor (*Have the evaluator(s)' experiences in program evaluation been reported?*) was not met in most of the research (n=22). Regarding

what partially means, to exemplify, if information about the stakeholders who are affected and who affect the evaluation is provided only about the participants (mostly only a group such as student-teachers) of the study, it is evaluated as partially. Those stakeholders might be anyone who is directly or indirectly involved in program and evaluation. When those stakeholders are not identified in the study, it will be not sufficient in terms of utility standard (Yarbrough et al., 2011).

To illustrate, in the context of utility standard, evaluators' experiences and expertises in program evaluation were not reported in the theses. Researchers specifically checked cover page, method and curriculum vitae sections for the analysis. However, experiences relevant to the evaluated program have been reported in T1 and T4. (please see Appendix for the sample of analysis). Following quotation is from T4 as an evidence of researcher's experiences in relation to the program evaluated: "In this context, in this study, first of all, the relevant literature was reviewed and possible problems related to the teaching practice course were identified based on the researcher's own experience" (p.57).

Another utility standard is about attention to stakeholders. Nearly half of the evaluations included different stakeholders as data sources. In Turkey, most of the evaluation studies have a reflective nature based on the reflections of various stakeholders like teachers, students, parents (Özdemir, 2009). Following quotation from A6 illustrates the case: "Data were collected through an adapted version of the [...] given to student teachers and graduates, focus group interviews with student teachers, and interviews with supervisors" (p. 251). However, the majority of the evaluations do not directly address the stakeholders who may affect and may be affected from the evaluation by only reporting the research participants. Some of the stakeholders seem to affect the evaluation in terms of developing data collection tools or data analysis. One of them identifying the stakeholders who may affect the evaluation is illustrated in A1 as follows: "Expert opinions of four academics from the curriculum and instruction department and three academics from elementary education department were obtained regarding whether the items in the item pool prepared could evaluate the pre-service elementary teacher education program or not; and expert opinions of two academics in the field of Turkish language were taken regarding the clarity of the items." (p. 832).

As another example from the utility standard can be negotiated purposes. Few evaluations reported about informing participants about the purpose of evaluation. To illustrate, a quotation from A14 is as follows: "The study group consisting of instructors was infromed about the purpose of the interview at the beginning of the interview and provided the program in use for their review." (p. 327)

Do ITE program evaluation research conducted in Türkiye in the last decade meet feasibility standards?

Results of the analysis of the evaluation research for the feasibility standards were presented in Table 3.

Table 3. Results on feasibility standards

Standard	Items/Questions	Yes	No	Partially
Feasibility	1. Has the planning information on the evaluation process been provided?	4	20	0
	2. Has it been reported how the evaluation process has been carried out?	6	10	8
	3. Have the participants' experiences and opinions relevant to the evaluated program been reported?	23	0	1
	4. Have independent (external) expert opinions been taken in the evaluation process?	11	2	11
	5. Have the resources utilized in the evaluation (such as human resources, equipment, training, travel) been reported?	2	22	0
	6. Has the cost of evaluation been reported?	0	24	0
	Total **f**	**46**	**78**	**20**
	%	**32%**	**54%**	**14%**

Table 3 showed that the ITE program evaluation research met the feasibility standard at the rate of 32 percent and partially met the relevant standard at the rate of 14 percent. However, the program evaluation research did not meet the feasibility standard at the rate of 54 percent. Whilst the descriptor (*Have the participants' experiences and opinions relevant to the program to be evaluated been reported?*) was met in most of the evaluation research (n=23), it was found that the descriptor (*Has the cost of evaluation been reported?*) was not met in any evaluation research (n=24). To exemplify what partially means, if independent expert opinions are obtained only in developing data collection instruments, it is evaluated as partially (not sufficient). It is expected to take independent experts' opinions throughout the whole evaluation process in terms of feasibility standards (Yarbrough et al., 2011).

One of the feasibility standards is project management. Few evaluations had information about planning the evaluation process. To illustrate, a quotation from A18 is as follows: "Data were collected in line with the evaluation plan." (p. 67). (A table is presented on page 67). Another feasibility standard is related to resource use. Resource use was not reported in most evaluations. Researchers specifically checked the evaluations to find out any signs of resource use. A quotation from one of the two evaluations (A3) reported about the resources is as follows: "For the collection of quantitative data, the first researcher traveled to the faculties of education in the sample group in line with the permission obtained from the universities and the developed questionnaire form was applied to pre-service teachers." (p. 149).

Do ITE program evaluation research conducted in Türkiye in the last decade meet accuracy standards?

Results on whether ITE program evaluation research met the accuracy standards were shown in Table 4.

Table 4. Results on accuracy standards

Standard	Items/Questions	Yes	No	Partially
Accuracy	1. Has the information on the evaluated program been provided?	13	4	7
	2. Has the information on the implementation context of the evaluated program been provided?	8	15	1
	3.Has the evaluation plan been reported to be organized based on needs?	4	18	2
	4. Has the evaluation plan been reported to be discussed with the relevant stakeholders?	0	24	0
	5. Has the evaluation been based on a program evaluation model?	14	10	0
	6. Has the program evaluation model been justified in the evaluation?	13	11	0
	7. Has the research method/design of the evaluation served the purpose of the evaluation?	20	0	4
	8. Has the research method/design of the evaluation been congruent with evaluation questions?	22	1	1
	9. Have the data collection instruments been introduced in the evaluation?	20	0	4
	10. Have the reasons why to use data collection instruments been explained in the evaluation?	15	7	2
	11. Has data triangulation (interview, observation, survey, etc.) been used in the evaluation?	15	6	3
	12. Have the data collection instruments used in the evaluation served the purpose of evaluation?	22	0	2
	13. Have the precautions required for increasing validity been taken in the evaluation?	13	6	5
	14. Have the precautions required for increasing reliability been taken in the evaluation?	12	4	8
	15. Has the path followed in the evaluation been explained?	8	13	3
	16. Have the data analysis techniques been justified in the evaluation?	14	10	0
	17. Have the roles of evaluator(s) been explicitly identified in the evaluation?	4	19	1
	18. Have the evaluation results been grounded on a theoretical basis?	12	12	0
	19. Have the evaluation results served the purpose of evaluation?	23	0	1
	20. Have the evaluation results been associated with the evaluation questions?	16	4	4
	21. Is the evaluation report accessible for all stakeholders?	24	0	0
Total	**f**	**292**	**164**	**48**
	%	**58%**	**32%**	**10%**

Table 4 displayed that the ITE program evaluation research were found to meet the accuracy standard at the rate of 58 percent and partially meet the relevant standard at the rate of 10 percent. However, the program evaluation research did not meet the accuracy standard at the rate of 32 percent. Whereas the descriptor (*Is the evaluation report accessible for all stakeholders?*) was completely met in all the evaluation research included in the meta-evaluation (n=24), it was observed that the descriptor (*Has the evaluation plan been reported to be discussed with the relevant stakeholders?*) was not met in any evaluation research (n=24). To exemplify what partially means, if the program to be evaluated is not explicitly described with its several dimensions, it is evaluated as partially. Details of the evaluated program will help better understand the purposes and procedures in the evaluation in terms of accuracy (Yarbrough et al. 2011).

One of the accuracy standards is about explicit program and context descriptions. Nearly half of the evaluations reported information about the evaluated program, yet, implementation context was not in nearly half of them. To illustrate, A5 reported information on both the evaluated program and implementation context. A quotation for the evaluated program is as follows:

> In Turkey, before 2006, CSP course was offered as an elective course by some universities in Turkey. In 2006, the Council of Higher Education updated teacher education programs, and from the academic year 2006–2007 CSP was included in the program and started to be implemented as a compulsory course. In 2011, the Council of Higher Education defined standards for the instruction of the course. The course content was described as follows […] (p. 347).

Another quotation from A5 on implementation context of the evaluated program is as follows:

> In the METU Faculty of Education, the course was put into practice in 2008. It was planned by four instructors from the faculty (from the departments of CEIT, ELE, and FLE). The course was intended to give pre-service teachers an opportunity to become involved in organizations serving the community in order to carry out tasks designed to contribute to a better society. According to the course objectives, at the end of the semester, students should be able to identify social issues related to education, and carry out voluntary tasks for organizations serving the community. [...] (p. 348)

Another example from accuracy standards can be about sound designs and analysis. Most of the evaluations utilized research methods that served the purpose of the study and to answer the evaluation questions. To illustrate, a quotation from A13 is as follows:

> In the study, concurrent triangulation strategy, which is one of the mixed method research types, was taken into consideration. Accordingly, in order to make a comprehensive analysis of the research problem, the researchers collected qualitative and quantitative data at the same time with both a quantitative and qualitative approach. Then, they combined these data and integrated them while interpreting the general results (Creswell, 2013). In line with Patton (2002), the quantitative approach used a pretest-posttest program evaluation design, while the qualitative approach used interviews and observation (p.50).

Do ITE program evaluation research conducted in Türkiye in the last decade meet propriety standards?

Results of the analysis of the evaluation research for the propriety standards were displayed in Table 5.

Table 5. Results on propriety standards

Standard	Items/Questions	Yes	No	Partially
Propriety	1.Has the information on permission from the ethics committee/commission to carry out the evaluation been provided?	1	23	0
	2. Has the information on rights and responsibilities of the participants been provided in the evaluation?	1	22	1
	3. Have the opinions of all participants been reported in the evaluation?	24	0	0
	4. Have recommendations based on the evaluation results been provided in the evaluation?	24	0	0
	5. Have the limitations of the evaluation been explained?	12	12	0
	6.Has it been explicitly reported whether there is conflict of interest in the evaluation?	0	24	0
	7. Has the information on financial matters such as income or expenditure been reported in the evaluation?	0	24	0
Total	**f**	**62**	**105**	**1**
	%	**37%**	**62%**	**1%**

As indicated in Table 5, the ITE program evaluation research were found to meet the propriety standard at the rate of 37 percent and partially meet the relevant standard at the rate of 1 percent. However, the program evaluation research did not meet the propriety standard at the rate of 62 percent. While the descriptors (*Have the opinions of all participants been reported in the evaluation?* and *Have recommendations based on the evaluation results been provided in the evaluation?*) were met in all the evaluation research (n=24), it was found that two descriptors (*Has it been explicitly reported whether there is conflict of interest in the evaluation?* and *Has the information on financial matters such as income or expenditure been reported in the evaluation?*) were not met in any evaluation research (n=24). To exemplify what partially means, if the participants are only asked permission for voice recording, it is evaluated as partially. In this case, they will not be completely informed about their rights and participants' knowledge of their rights and responsibilities makes the evaluation fair and legal (Yarbrough et al., 2011).

One of the propriety standards is about human rights and respect. However, the majority of the evaluations reported no ethical issues such as participants' rights and responsibilities or ethical permissions. To illustrate, one of the evaluations (T1) reported about getting ethical approval by the following quotation: "Since the participants would answer the questions in Turkish, the interview schedule was translated into Turkish with the informed consent form [...]) for the Ethics Committee Approval (METU Human Subjects Ethics Committee). " (p.80). The same evaluation study (T1) reported about informing participants about their rights and responsibilities by the following quotation:

> It is also important to get the consent of the interviewees' prior to the interview as well as in the beginning of the interview by repeating the reason why the data is collected, how the data will be used, what kind of questions will be asked and inform them about the possible benefits or risks of the research on part of the interviewee. In this research, the participants were given or sent information consent after they agreed to participate via by phone or personally or replied to the e-mail sent by the researcher as an invitation to participate in the research study. Before the interview, all the participants agreed to sign the consent form and their emails were taken in order to send them the transcribed data and check for their feedback. (p. 87)

Another propriety standard is about transparency and disclosure. Regarding explicitly reporting limitations of the evaluation with stakeholders, half of the evaluations had information about the limitations. To illustrate, a quotation from A6 is as follows:

> On a final note, it should be explicitly stated here that one of the limitations on the results drawn from this study came from the existing small number of the university supervisors responsible for the practicum programme. The comparatively fewer number of programme graduates who returned the online survey is another limitation, which might have been caused by the researchers' preference for reaching graduates from the classes of the near past. The greater number of open-ended survey items given to this group might have also resulted in reluctance to participate. (pp. 269-270)

Do ITE program evaluation research conducted in Türkiye in the last decade meet accountability standards?

Results on whether ITE program evaluation research met the accountability standards were presented in Table 6.

Table 6. Results on accountability standards

Standard	Items/Questions		Yes	No	Partially
Accountability	1. Has the evaluation been described in detail with all dimensions such as planning, procedures and outcomes?		3	1	20
	2. Have the program evaluation standards been reported in the evaluation process?		0	24	0
	3. Has it been reported that the evaluation process has been conducted in collaboration with program stakeholders?		1	2	21
	Total	**f**	**4**	**27**	**41**
		%	**6%**	**38%**	**57%**

Table 6 showed that the ITE program evaluation research met the evaluation accountability standard at the rate of 6 percent and partially met the relevant standard at the rate of 57 percent. However, the program evaluation research did not meet the evaluation accountability standard at the rate of 38 percent. Whereas two descriptors (*Has the evaluation been described in detail with all dimensions such as planning, procedures and outcomes?* and *Has it been reported that the evaluation process has been conducted in collaboration with program stakeholders*?) were partially met in most of the evaluation research (n=20, n=21), it was found that the descriptor (*Have the program evaluation standards been reported in the evaluation process?*) was not met in any evaluation research (n=24). To exemplify what partially means, if the evaluator only works in collaboration with participants rather than other program stakeholders, it is evaluated as partially meeting the standard. It is expected to make collaboration with program stakeholders in several terms such as taking expert opinions, getting permissions to carry out the evaluation and so on in terms of accountability standards (Yarbrough et al., 2011).

One of the evaluation accountability standards is about evaluation documentation, which means detailed documentation of evaluations from planning to obtaining outcomes. Few evaluations had detailed information on the evaluation process whilst the majority of them had partial information such as only documenting results without detailed information on planning or procedures. To exemplify, A18 reported about planning and procedures under method section and outcomes under results section of evaluation report.

Discussion, Conclusion and Suggestions

The current meta-evaluation study investigated to what extent ITE program evaluation research conducted in the last decade in Türkiye complied to program evaluation standards set by JCSEE by utilizing the rubric developed by the researchers based on the relevant standards. The rubric was developed to assess evaluation reports in terms of utility, accuracy, feasibility, propriety and accountability standards by following the steps such as compiling an item pool based on JCSEE standards, taking expert opinions, piloting the rubric with evaluation research, eliminating or rewriting items, ordering the items and finalizing the rubric. The Program Evaluations Meta-evaluation Checklist developed by Stufflebeam (2012) was frequently used in literature; however, a new instrument was needed to serve the purpose of assessing written program evaluation reports since all items in the checklist developed by Stufflebeam (2012) would not be completely applicable for assessing the written reports. Moreover, the study conducted by Tingle et al. (2003) employed an instrument based on only accuracy standards to assess evaluation research. Therefore, the meta-evaluation rubric developed by the researchers in the current study was considered to be a practical and comprehensive instrument to serve the technical meta-evaluative purposes. Thus, researchers can utilize the rubric for conducting program evaluations based on program evaluation standards to increase the quality of their evaluations and meta-evaluators can use it for conducting meta evaluation studies.

The results of meta-evaluation of ITE program evaluations demonstrated that the research mostly met accuracy standards at the rate of 58 percent whilst propriety and feasibility standards were not met at the rate of 62 percent and 54 percent, respectively. However, the meta-evaluation conducted by Widmer (2000) based on JCSEE found that the evaluation research met utility, feasibility and propriety standards strongly; on the other hand, accuracy standard was not met sufficiently in terms of purposes, procedures, valid and reliable measurement, and justified conclusions. Firstly, the difference in results regarding accuracy standard might result from the fact that Widmer (2000) examined non-educational evaluations in various fields like social policies or environment. Those kinds of evaluations might differentiate from educational evaluations in terms of how the research paper was structured. The ITE program evaluation research included in the current study were all structured in terms of purpose, method, results and conclusions, which in turn might increase the possibility of meeting the accuracy standard. Secondly, the difference in findings regarding propriety and feasibility standards might depend on the way the evaluations were assessed. To explain, Widmer (2000) assessed the evaluations based on the JCSEE standards through both written materials and interviews with evaluators and stakeholders. The current meta-evaluation, on the other hand, was designed based solely on written materials; therefore, if the indicators of standards were not explicitly reported in the written evaluation research, it would be interpreted as the lack of the relevant indicator. Social interactions between meta-evaluator and program stakeholders are of importance for mutual understanding of meta-evaluation purposes and results (Stufflebeam 2001). Interviews with program stakeholders might facilitate understanding the compliance of evaluations with propriety and feasibility standards in Widmer's (2000) meta-evaluation.

Accuracy standard in the current study was found to be the strongest in methodological terms such as research design. Similarly, Akıncı and Köse (2022) concluded that accuracy standard was the standard that was met at the highest level in evaluations of teacher training programs. Also, the study conducted by Tingle et al. (2003) demonstrated that evaluation research on school-based smoking prevention programs met accuracy standards mostly in terms of research design. However, Scott-Little et al. (2002) found in their meta-evaluation that after-school program evaluation reports were weak in terms of research designs. Yüksel and Akın (2013) also reported in their meta-evaluation that justified conclusions as the indicator of accuracy standard did not suffice in Student Achievement Examination reports. Those different findings might be the result of the design of evaluations. The current study assessed evaluation research on ITE programs, and they were more methodologically structured. This may have an impact upon the reason for methodological strength of the evaluations. On the other hand, 10 out of 24 evaluation studies were not based on a program evaluation model, which is also reported as a problematic issue in program evaluation studies conducted in our country (e.g. Akıncı ve Köse, 2021; Kurt & Erdoğan, 2015; Kürüm-Yapıcıoğlu, Atik-Kara, & Sever, 2016). Program evaluation models demonstrate which ways to follow and which types of evaluation to implement (Oliva & Gordon, 2013). Utilizing a model when conducting program evaluation study is a significant indicator in accuracy standard since use of an evaluation model makes the evaluations more systematic. Therefore, program evaluation studies should be strengthened by drawing on evaluation models to be more accurate.

Feasibility standard was found to be the standard that was met slightly in the current study. Information about planning of evaluation, utilizing resources and evaluation costs were the indicators that most research lacked. However, Yasar et al. (2005) reported that evaluations of elementary education teacher training programs were sufficient in feasibility. The meta-evaluation conducted by Akıncı and Köse (2020) on initial teacher education programs demonstrated that the feasibility standard was moderately met. Almost all of the examined studies, except one partial, met the reporting experiences and opinions of the relevant program evaluated. Although the research assessed the evaluations based on written evaluation research, those different results may depend on several factors such as which indicators were selected as references for each standard, whether the study was a research article or thesis/dissertation or how the meta-evaluators rated the indicators for each study. To exemplify, detailed information could be found in theses or dissertations in comparison with research articles, which in turn may affect rating the standard indicators by meta-evaluators. Here, as a suggestion to researchers, information about how to ensure effectiveness and efficiency of evaluations should be depicted more systematically (including expert opinions as a way to ensure validity and reliability and the resources utilized and cost of the evaluation), either in the form of figures or narrations.

The results of the current meta-evaluation study indicated that utility standard was met in evaluation research at the level of 46 percent, which could be interpreted as a relatively sufficient level. Yasar et al. (2005) found in their meta-evaluation that teacher training program evaluations were partly sufficient in utility standards. Similarly, Akıncı and Köse (2020) reported that utility standards were moderately met in teacher training program evaluations. Those results may be considered consistent with each other. Utility standard is about stakeholders in general (Yarbrough et al., 2011). To meet the utility standard in program evaluation studies, evaluators (researchers) should be more responsive to the stakeholders' expectations, involvement and contributions, and they should provide more detailed information about the stakeholders in their evaluation reports. In addition to that, giving a place to evaluator(s)' expertise and experiences in program evaluation would strengthen the studies through the utility lenses.

Lastly, the current meta-evaluation showed that the accountability standard was partially met in evaluations of initial teacher education programs. Studies that addressed accountability standards were not found in the literature. The accountability standard was added to JCPES in 2011 (Yarbrough et al., 2011). Akıncı and Köse (2020, 2022) reported that they did not include accountability standards in the checklist they developed to conduct meta-evaluation since accountability is directly related to meta-evaluation. However, the current meta-evaluation study focused on accountability standards as well as the other standards to understand whether the written research on initial teacher education program evaluation met the detailed description, compliance with program evaluation standards and collaboration with program stakeholders, which are the indicators of evaluation accountability standards. Further evaluation studies might include accountability standards as a way to strengthen their research considering documentation and internal and external meta-evaluation processes.

To conclude, the current study revealed that initial teacher education program evaluation research in Turkish context were not sufficiently based on program evaluation standards. The evaluations need to be improved in terms of feasibility, propriety, utility and accountability standards. Even, accuracy standard is expected to be met at higher rates. To this end, evaluators can utilize the meta-evaluation rubric developed in this study to conduct their evaluation studies and to write their evaluation reports. Thus, the meta-evaluation rubric may be a roadmap for them to plan, carry out and report the evaluation process and results.

The study has some limitations. Firstly, the results of the current meta-evaluation are limited to the data obtained from criterion-based selected research on ITE program evaluations published between 2010-2020 in Türkiye. Secondly, the results are limited to the analysis from the rubric developed by the researchers. Last, the researchers of this meta-evaluation conducted evaluation of ITE program evaluation studies. External evaluators may be involved in the meta-evaluation to increase reliability.

As for recommendations for the future research studies, different program evaluations such as K-12 programs or teacher professional development programs to identify merit and worth of those evaluations by utilizing the rubric developed in this study might be conducted. Furthermore, to fully understand to what extent program evaluations meet the program evaluation standards, interviews with program evaluators and stakeholders would be valuable to support the data obtained from use of a rubric. Varying data collection tools helps increase reliability of the research and provide comprehensive data. Future studies can be conducted in different countries with an aim to make cross cultural evaluation comparisons.

References

Akıncı, M., & Köse, E. (2020). Türkiye'de öğretmen yetiştiren programlara ilişkin bir meta-değerlendirme çalışması. In *Uluslararası Pegem Eğitim Kongresi* (pp. 174-175). Diyarbakır: Dicle Üniversitesi.

Akıncı, M., & Köse, E. (2021). Research trends of program evaluation studies conducted between 2010-2019 in Turkey. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 50*(1), 77-120. doi:10.14812/cufej.688142

Akıncı, M., & Köse, E. (2022). A meta-evaluation research on teacher training programs in Türkiye. *International Journal of Progressive Education, 18*(4), 209-222. doi:10.29329/ijpe.2022.459.15

Bowen, G. A. (2009). Document analysis as a qualitative research method. Qu*alitative Research Journal, 9*(2), 27-40. doi:10.3316/QRJ0902027

Cook, T. D., & Gruder, C. L. (1978). Meta-evaluation research. *Evaluation Quarterly, 2*(1), 5-51.

Cooksy, L. J., & Caracelli, V. J. (2009). Meta-evaluation in practice: Selection and application of criteria. *Journal of MultiDisciplinary Evaluation, 6*(11), 1-15.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston: Pearson Education, Inc.

Gözütok, D. (2006). Program değerlendirme. In M. Gültekin (Ed.), *Öğretimde planlama ve değerlendirme* (pp. 175-190). Eskişehir: Anadolu Üniversitesi Yayınları.

Hedler, H. C., & Gibram, N. (2009). The contribution of metaevaluation to program evaluation: Proposition of a model. *Journal of MultiDisciplinary Evaluation, 6*(12), 210-223.

Kablan, Z. (2011). Analysis of elementary mathematics curriculum evaluation studies. *Elementary Education Online, 10*(3), 1160-1177.

Kurt, A., & Erdoğan, M. (2015). Content analysis and trends in curriculum evaluation research: 2004-2013. *Education and Science, 40*(178), 199-224.

Kürüm-Yapıcıoğlu, D., Atik-Kara, D., & Sever, D. (2016). Türkiye'de program değerlendirme çalışmalarında eğilimler ve sorunlar: Alan uzmanlarının gözüyle. *Uluslararası Eğitim Programları ve Öğretim Çalışmaları Dergisi, 6*(12), 91-113.

Lincoln, Y. S., & Guba, E. G. (1980). The distinction between merit and worth in evaluation. *Educational Evaluation and Policy Analysis, 2*(4), 61-71.

Melrose, M. (1998). Exploring paradigms of curriculum evaluation and concepts of quality. *Quality in Higher Education, 4*(1), 37-43. doi:10.1080/1353832980040105

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation*, 7, 1-6. doi:10.7275/q7rm-gg74

Oliva, P. F., & Gordon, W. R. (2013). *Developing the curriculum* (8th ed.). New Jersey: Pearson Education, Inc.

Ornstein, A. C., & Hunkins, F. P. (2017). *Curriculum: Foundations, principles and issues* (7th ed.). Essex: Pearson Education Limited.

Özdemir, S. M. (2009). Eğitimde program değerlendirme ve Türkiye'de eğitim programlarını değerlendirme çalışmalarının incelenmesi. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 2*, 126-149.

Patton, M. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Sağlam, M., & Yüksel, İ. (2007). Program değerlendirmede meta-analiz ve meta-değerlendirme yöntemleri. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi,* 18, 175-187. Retrieved from https://dergipark.org.tr/tr/download/article-file/55419

Sanders, J. R. (1994). *The program evaluation standards. How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.

Scott-Little, C., Hamann, M. S., & Jurs, S. G. (2002). Evaluations of after-school programs: A meta-evaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation, 23*(4), 387-419.

Scriven, M. (1969). An introduction to meta-evaluation. *Educational Products Report, 2*, 36-38.

Stufflebeam, D. L. (1978). Meta evaluation: An overview. *Evaluation and The Health Professions*, *1*(1), 17-43.

Stufflebeam, D. L. (2000a). The methodology of metaevaluation as reflected in metaevaluations by the Western Michigan University Evaluation Center. *Journal of Personnel Evaluation in Education, 14*(1), 95-125.

Stufflebeam, D. L. (2000b). The methodology of metaevaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models* (pp. 457-471). Boston: Kluwer Academic Publishers.

Stufflebeam, D. L. (2001). The metaevaluation imperative. *American Journal of Evaluation, 22*(2), 183-209.

Stufflebeam, D. L. (2002). Foundational models for 21st century program evaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 33-85). New York: Kluwer Academic Publishers.

Stufflebeam, D. L. (2004). A note on the purposes, development, and applicability of the Joint Committee Evaluation Standards. *American Journal of Evaluation, 25*(1), 99-102.

Stufflebeam, D. L. (2011). Meta-evaluation. *Journal of MultiDisciplinary Evaluation, 7*(15), 99-158.

Stufflebeam, D. L. (2012). Program evaluations meta-evaluation checklist (Based on The Program Evaluation Standards). Retrieved from https://usaidlearninglab.org/sites/default/files/resource/files/mod3_meta_evaluation_checklist_1.pdf

Tingle, L. R., DeSimone, M., & Covington, B. (2003). A meta-evaluation of 11 school-based smoking prevention programs. *Journal of School Health, 73*(2), 64-67.

Widmer, T. (2000). Evaluating evaluations: Does the Swiss practice live up to the 'program evaluation standards'?. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 67-80). Kalamazoo: Western Michigan University, Evaluation Center.

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *Journal of Effective Teaching*, *7*(1), 3-14.

Yağan, S. A. (2019). Program değerlendirme alanında yayınlanmış doktora tezlerinin meta değerlendirmesi (2015-2018). *Elektronik Eğitim Bilimleri Dergisi, 8*(16), 188-208.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Yasar, S., Gultekin, M., Kose, N., Girmen P., & Anagun, S. (2005). The meta-evaluation of teacher training programs for elementary education in Turkey. In *Teacher Education: Local and Global Proceedings of the 33rd Annual Australian Teacher Education Association Conference* (pp. 498-504). Australia.

Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri.* Ankara: Seçkin.

Yüksel, İ., & Akın, Z. (2013). Öğrenci Başarısını Belirleme Sınavının (ÖBBS) öğrenci değerlendirme standartları kapsamında değerlendirilmesi: Bir meta değerlendirme araştırması. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 32*(1), 473-495.

Yüksel, İ., & Sağlam, M. (2011). *Türkiye için program değerlendirme standartları oluşturma çalışması*. 1. Uluslararası Eğitim programları ve Öğretim Kongresi, Eskişehir.

Appendix

ARTICLE/THESIS: A5

Standard	Items/Questions	Yes	No	Partially	Notes / Evidence (Please provide details for your response, add evidence from the studies)
1. (Utility)	1. Have the evaluator(s)' expertise in program evaluation been reported?	X			The evaluator(s) was stated to work in Curriculum and Instruction Department.
	2. Have the evaluator(s)' experiences in program evaluation been reported?		X		
	3. Have the evaluator(s)' experiences relevant to the evaluated program been reported?		X		
	4. Have the stakeholders who may be affected by the evaluation (such as experts, participants) been identified?			X	Only participants (pre-service teachers and teacher educators)
	5. Have the stakeholders who may affect the evaluation (such as experts, participants) been identified in the study?	X			Experts and participants (pre-service teachers and teacher educators)
	6. Have the participants been informed about the purpose of the evaluation?		X		Not reported
	7. Have the evaluation data been consistent with the purpose of evaluation?	X			
	8. Have the evaluation data been collected from different sources (such as teachers, parents, school principals)?	X			Pre-service teachers and teacher educators
	9. Has the evaluation been designed to meet the needs of stakeholders?		X		Different stakeholders were stated, yet needs analysis was not reported.
	10. Has the purpose of the evaluation been reported explicitly and clearly?	X			
	11. Has the evaluation process been reported explicitly and clearly?	X			Observations, in particular, were reported in detail; what was done for input, process, product and context was reported in detail.
	12. Have the evaluation results been reported explicitly and clearly?	X			
	13. Have the evaluation results been reported in a meaningful way for the stakeholders (in a non-technical language)?	X			
	14. Has the communication between evaluator(s) and stakeholders been reported in the evaluation process?	X			By means of prolonged engagement, researcher spent enough time in research site…p. 353
	15. Have the potential benefits or effects of evaluation results for stakeholders been reported?	X			
	16. Has the evaluation been justified in the relevant conditions and context?	X			p. 348

2. (Feasibility)	17. Has the planning information on the evaluation process been provided?		X		The plan was not reported directly.
	18. Has it been reported how the evaluation process has been carried out?	X			It was reported by associating with researcher roles. p.351-2-3
	19. Have the participants' experiences and opinions relevant to the evaluated program been reported?	X			
	20. Have independent (external) expert opinions been taken in the evaluation process?			X	Expert opinions were obtained for each data collection tool.
	21. Have the resources utilized in the evaluation (such as human resources, equipment, training, travel) been reported?		X		
	22. Has the cost of evaluation been reported?		X		
3. (Accuracy)	23. Has the information on the evaluated program been provided?	X			p. 348
	24. Has the information on the implementation context of the evaluated program been provided?	X			p. 348
	25.Has the evaluation plan been reported to be organized based on needs?		X		
	26. Has the evaluation plan been reported to be discussed with the relevant stakeholders?		X		
	27. Has the evaluation been based on a program evaluation model?	X			CIPP
	28. Has the program evaluation model been justified in the evaluation?	X			
	29. Has the research method/design of the evaluation served the purpose of the evaluation?	X			Case study
	30. Has the research method/design of the evaluation been congruent with evaluation questions?	X			
	31. Have the data collection instruments been introduced in the evaluation?	X			
	32. Have the reasons why to use data collection instruments been explained in the evaluation?	X			
	33. Has data triangulation (interview, observation, survey, etc.) been used in the evaluation?	X			Questionnaires, interview protocols, observations and document analysis were utilized in context, input, process and product evaluations.
	34. Have the data collection instruments used in the evaluation served the purpose of evaluation?	X			
	35. Have the precautions required for increasing validity been taken in the evaluation?	X			It was reported in detail p.353
	36. Have the precautions required for increasing reliability been taken in the evaluation?	X			It was reported in detail p.353
	37. Has the path followed in the evaluation been explained?	X			
	38. Have the data analysis techniques been justified in the evaluation?	X			
	39. Have the roles of evaluator(s) been explicitly identified in the evaluation?	X			Prolonged engagement
	40. Have the evaluation results been grounded on a theoretical basis?	X			CIPP
	41. Have the evaluation results served the purpose of evaluation?	X			
	42. Have the evaluation results been associated with the evaluation questions?	X			Context, input, process and product evaluations
	43. Is the evaluation report accessible for all stakeholders?	X			

4. (Propriety)	44.Has the information on permission from the ethics committee/commission to carry out the evaluation been provided?		X		
	45. Has the information on rights and responsibilities of the participants been provided in the evaluation?		X		
	46. Have the opinions of all participants been reported in the evaluation?	X			
	47. Have recommendations based on the evaluation results been provided in the evaluation?	X			
	48. Have the limitations of the evaluation been explained?	X			
	49.Has it been explicitly reported whether there is conflict of interest in the evaluation?		X		
	50. Has the information on financial matters such as income or expenditure been reported in the evaluation?		X		
5. (Evaluatoin accountability)	51. Has the evaluation been described in detail with all dimensions such as planning, procedures and outcomes?			X	Planning was not reported explicitly, however, procedures and outcomes were reported in detail.
	52. Have the program evaluation standards been reported in the evaluation process?		X		
	53. Has it been reported that the evaluation process has been conducted in collaboration with program stakeholders?			X	Only participants