Using a CIPP-Based Model for Evaluation of Teacher Training Programs in a Private-sector EFL Institutes

Maryam Khaksar[1], Gholam Reza Kiany[2]*, Parvaneh ShayesteFar[3]

[1,2]English Language Teaching Department; Tarbiat Modares University, Tehran, Iran
[3]English Language Teaching Department, Farhangian Teacher Education University, Tehran, Iran

Abstract

As teachers play a significant role in language learners' academic achievement, their training programs should be subjected to ongoing evaluation and analysis in order to ascertain teachers are equipped with adequate knowledge and skills. This study aimed to evaluate a private sector pre-service teacher-training program using the Context, Input, Process and Product (CIPP) model. For such a purpose, the program policies, planned course of actions and real practices, as well as the outcomes were closely studied through a sequential mixed-method design. The required data were obtained from the program's stakeholder layers (N=296: 100 supervisors, 58 trainers, 100 teachers, and 38 trainees) participating in data collection phases using multiple methods and instruments including interviews, observations and surveys. All instruments were subjected to detailed validation and reliability investigations. The results revealed degrees of positive perspectives towards the program's functioning and outcomes in training reflective, creative and energetic teachers. Although the participants reported optimistic views about the program quality, required modifications and essential improvements were strongly stressed by the trainers in terms of the length of the course, focus on classroom management, and provision of practice opportunities. The findings have implications for evaluation of teacher training programs implemented in similar contexts elsewhere.

Keywords: *Program Evaluation, Teacher Training Program, CIPP Model of Evaluation, Teacher Qualifications, Program Quality*

Introduction

Teacher education programs have been found to have a tremendous influence on teachers' performance and consequently on students' achievements (Crandall, 2000). Thus, teacher education programs and the elements affecting them have been extensively studied over the past two decades (Berg et al., 2023; Clark & Newberry, 2019; Dursun et al., 2023; Ellis &

Childs, 2019; Farrell, 2011, 2012; Grossman, 2021; Harris & Sass, 2011; Loewenberg Ball et al., 2008; Ping et al., 2018; von Hippel et al., 2016). These studies have revealed a number of components affecting the quality of teacher education programs. Among these elements, *context*, *materials* and the *methods* through which a program is delivered are argued as the key elements (Freeman & Johnson, 1998). However, the program's quality is not assured only with the existence these components. The way each function interacts with others, and aligns with the program objectives is worth evaluating (Levine, Mitoma, Anagnostopoulos, & Roselle, 2022). The need for evaluation of teacher education programs from this perspective has consequently led to the application of program evaluation models to ensure program's effectiveness and success (Barry et al., 2020; Darling-Hammond & Bransford, 2007).

The chosen evaluation model must consider the context of the program, the materials used, the delivery of said material, and the results (McNamara, 2008) , which creates a need for formative and summative assessment techniques. (Frye & Hemmer, 2012) . This need would give an advantage to more inclusive evaluation models such as the Kirkpatrick (1996) Model or Kaufman (1994) Model or Stufflebeam's (2003) Context, Input, Process, and Product (CIPP) evaluation model. From another standpoint, considering the classification of evaluation models offered by Worthen et al. (1997) the requirements of teacher-training program evaluation put it in the category of management-oriented evaluation models (young Lee et al., 2019). In this model, all aspects of the program are evaluated and the results are used by decision makers to make necessary modifications (Stufflebeam, 2004).

The CIPP model was developed to ensure quality of programs and diagnose issues. Hence, it is widely used in educational settings. This model allows the evaluator to make sets of criteria that best fit the program they are evaluating for each phase and use them accordingly (Darussalam, 2010). Considering the abovementioned criteria, the study employed Stufflebeam's (2003) CIPP model. The model seems to be "a comprehensive framework for conducting formative and summative evaluations of projects, personnel, products, organizations, and evaluation systems" (Stufflebeam & Shinkfield, 2007, p. 325). The sources of questions addressed by this model are the involved stakeholders and the evaluator's views on the questions that must address a program's values. In brief, in Stufflebeam and Shinkfield's (2007) view, *context* evaluation provides information about what the needs of the trainees (student-teachers in the present case) entering the program are, what the expectations and perspectives of their instructors or other stakeholders are, and what assets and funding opportunities could be used to address the targeted needs.

Input evaluation focuses on the service strategies and the budget planned to address the needs of the context. They also discuss that evaluation of *process* aims to ascertain if the planned strategies are implemented. Finally, the *product* evaluation examines the intended and unintended outcomes of the program to judge the program's effectiveness, achievement, and sustainability.

Following Foroozandeh (2008) who evaluated Teaching English as a Foreign Language (TEFL) program at Master Level in Iran using the CIPP model, it was found that this model could serve as a good basis for evaluating teacher education programs. The present evaluation framework was therefore adopted to obtain a clear picture of the objectives of a non-governmental English Language Teaching (ELT) training program. Despite the growing

interest in employing CIPP model around the world (e.g., Zhang et al., 2011; Stufflebeam, 2003, 2007), there is a dearth of such studies in Iranian context.

This study aims to evaluate the teacher-training course of one of the largest language institutions in Iran, Safir Language Academy. The reason why this particular institution was chosen for this study is that language education in this country relies heavily on the private sector (Iranmehr & Davari, 2018), and this particular institution is the second largest language-learning establishment in the country with over sixty branches nationwide (Safir-Language-Academy, 2023), which would make for a reasonable sample size.

Teacher education programs have often been studied in the context of universities, for instance Masoumpanah et al. (2017) have evaluated the English teacher training program at Teacher Education University which is a preparatory university for prospective teachers (Ghasemi et al., 2020). Despite the importance of such evaluative work, the quality and sustainably of the work done in the private sector remains a question. Given the private sector's prominent role in ELT, it is crucial to gain a better understanding of their teacher education programs. Among the institutes operating in Iran, Safir claims to have one of the most unified teacher training programs following the framework of Communicate Language Teaching (CLT) (Safir-Language-Academy, 2022). The program is delivered by trainers all of whom have advanced degrees in ELT (Safir-Language-Academy, 2022). The quality of the courses is constantly under supervision and all trainers are required to follow similar principles designed by the Teacher Training Department (TRD). The graduates are later observed using the standards taught during the course, enabling the researchers to test the sustainability of those standards. Unity of the materials and trainers' instructions and the possibility of systematized follow-up evaluation of the graduates make this course a valid candidate for evaluation.

Context of the Study

Safir Language Academy is a non-governmental institution, with its headquarters in the capital (Tehran) and around 70 district offices in different provinces of the country. Its TRD offers a four-week long program to prospective English teachers. They intend to incorporate standards CLT and the standards of Certificate in English Language Teaching to Adults (CELTA). It also includes the incorporation of some organizational values. The four-week training course comprises modules of lesson planning, teaching language systems, receptive and productive skills, and error treatment. The lessons are delivered using a comprehensive PowerPoint Presentation, worksheets and handouts. The program ends with a teaching demonstration, which defines whether the trainees are qualified to pass the course, or not.

Methods

Participants

According to the research design of this study, 296 participants comprising of 58 trainers, 100 teachers, 38 trainees and 100 supervisors participated in the different phases of the study.

Trainers: 10 teacher trainers, selected through purposive sampling, participated in this phase. Since an important phase of the study was the phase of instrument development, it was crucial that the respondents be experienced enough as trainers and fully aware of the policies made

and the rationale behind them. Hence, teacher trainers with teaching experience above 10 years were selected ($\bar{x}$=12.2 years).

*Trainees: A*s for the trainees, convenience sampling was carried out and two groups of trainees, (N=38) were selected. Table 1 shows their distribution.

Table 1

Distribution of Participants by University Degrees and Majors

Majors	Trainers		
	PhD	Master's	Bachelor's
ELT	4	23	1
Non-ELT	0	22	8
	Supervisors		
	PhD	Master's	Bachelor's
ELT	4	66	29
Non-ELT	0	10	11
	Teachers		
	PhD	Master's student/graduate	Bachelor's student/graduate
ELT	2	49	27
Non-ELT	0	6	16

Table 2

Distribution of Participants by Gender

	Male	Female
Trainers	11	47
Supervisors	40	80
Teachers	40	60

Instruments and Procedures

This study was conducted in four phases. Firstly, the evaluation scheme was developed through conducting interviews with trainers and developing a questionnaire. Secondly, the validity of the questionnaire was measured. The third phase included the analysis of the gathered data. In the fourth phase, the researchers observed two whole courses and the trainees were asked to keep journals of their learning experiences after each session in order to triangulate the data.

Evaluation Scheme Development Phase

To design the instrument (henceforth called CIPPTTCP), the first step was to explore the literature. The analysis revealed a number of standards expected of competent teachers and teacher education programs. This led to the development of interview questions (see Appendix A). Interviews with 10 teacher trainers were conducted from May to August 2018. Scheduled appointments were set to conduct one-to-one recorded interviews with each individual trainer at Safir. The interview questions were constructed in line with the existing literature on teacher-training programs' quality, process, effectiveness and goal-achievement. The interview guide included seven questions seeking how teacher educators felt about these. The questions were given to these respondents prior to the interview sessions. Trainers consented to being recorded

while talking. Each interview lasted between 30 to 90 minutes. The researcher also took notes and asked follow-up questions where needed. Before conducting the interviews, the content and face validity of the interview questions were reviewed by two experienced teacher trainers from two state teacher-education universities in Iran. Necessary modifications were made after of these viewings. Interview data were significantly valuable since they provided the researcher with a thick data stream on CIPP components. Recordings were transcribed and coded using MaxQDA. The process was repeated until no new codes were found. The analysis of each interview was done in two phases: one in which the interview was analyzed separately, regardless of other interviews. This stage of analysis is referred to as "vertical analysis" (Miles et al., 1994). After, the interviews were analyzed using "constant comparison analysis" (Glaser & Strauss, 1967) wherein the identified codes in each interview are consistently checked in order to weed out similarities and differences in codes and patterns. The coding system used in this study followed the sequence of open coding, axial coding and selective coding process (Strauss & Corbin, 1997), as tools for identifying thematic categories. Hence, the researcher began the content analysis process by exploring the interviews and segmenting them into independent codes/units, each with a distinct theme. After, these units were organized into condensed units that were then classified under their relevant categories. Therefore, the initial list included 489 codes/units. They were then analyzed further through axial coding. This was done in order to categorize the relevant codes/units, and form categories under which the related codes could be grouped (Strauss & Corbin, 1997). After this stage, the number of codes was reduced to 157. Glaser and Holton (2004) recommend that researchers delimit their search for codes to the core variable(s), accordingly, the codes/units selected in the last stage of coding were those relevant to the core variables of the current study, i.e. context (36 items), input (35 items), process (43 items) and product (43 items). To this end, the list was analyzed once more through selective coding, and the themes created in the previous stage were further scrutinized. As a result, unnecessary themes were left out leaving 138 relevant themes. The following extract from a trainer might help clarify the way coding was done:

Excerpt #1:
"One of the most important features that a person needs to have or needs to gain in the course if to pay attention, to care about how his or her students feel." (trainer #5, July, 2018)

The meaning unit "supportive/encouraging" was extracted. This condensed meaning unit was later classed under the category: "context".

After this phase, an instrument was developed using 138 themes extracted from the analysis of interviews, the available checklists, particularly Stufflebeam's CIPP checklists, and the relevant views found in the existing literature (Celce-Murcia & McIntosh, 1991; Chambless, 2012; Durik & Harackiewicz, 2007; Schacter & Thum, 2004). These data sources were employed for the development of the first part of CIPPTTCP questionnaire. The items were designed to elicit data on the quality, effectiveness and impact of the program. Since the trainers are those best familiar with the course, the researcher tried to obtain as many of their opinions as possible. Consequently, consecutive sampling also known as total enumerative sampling

(Daniel, 2011), was used. The trainers were aged 27 to 63. Branch supervisors can be considered as the end users of the program because they are the ones who will eventually work with graduates of the program. 120 supervisors, aged 24-52 were randomly chosen and asked to fill out the scale. As trainees sit the course and receive the lessons, is important to consider their point of view. Thus, 100 of the recently graduated ones, aged 18 to 35, were randomly selected. The scales were uploaded online and the link was sent to the participants using emails or text massages. Data collection process took about one month. After administrating the scale, 37 out of 45 trainers, 111 out of the 120 supervisors, and 42 out of the 100 graduates had completely responded to the instrument; therefore, the rest were removed from the final data set.

Validation Phase

To check the content and face validity of the instrument, three ELT university professors were asked to analyze the instrument. Additionally, three ELT-educated trainers were also asked to evaluate the questionnaire to make sure its content aligns with the institute's standards. Then, CIPPTTCPs were administered among 265 respondents and 190 were returned (a response rate of 71.6%). To examine the factor structure of the questionnaire, the validation involved a series of Factor Analyses, both Exploratory and Confirmatory Factor Analysis (CFA). First, to examine the internal structure of the set of the scales constructed, Exploratory Factor Analysis (EFA) was run for all scales, as follows.

Regarding the first section of the CIPPTTCP, the findings showed that 22 factors with Eigenvalues greater than one could be extracted, explaining 78% of the total variance. An examination of the content of the items of this scale provided empirical structure for the existence of 4 subscales. It is noteworthy that while these 4 factors emerged in EFA, Confirmatory Factor Analysis (CFA) with AMOS (version 18) was carried out to determine the adequacy of the factor loadings and more information about the structural measurement. The results of exploratory factor analysis showed that out of 149 items 85 items were kept. While these 4 factors emerged in EFA, Confirmatory Factor Analysis CFA with AMOS (version 18) was carried out to determine the adequacy of the factor loadings and the standardized residuals, and more information about the structural measurement.

In the present CFA-AMOS run, the normed Chi-square (shown by CMIN/DF) showed the value of ≤5 that according to Marsh and Hocevar (1985) indicates a reasonable model fit. However, CIMN/DF<3 indicates the best fit between hypothetical model and sample data (Kline, 1998). The inspection of values of the normed chi-square (CMIN/DF) and other Goodness-of-Fit Indices showed that the modified model indicates a reasonably fit structured model, with $\chi2=3479$ and CMIN/DF=3.7. Although the Goodness-of-Fit Index (GFI) and Comparative Fit Index (CFI) did not meet the recommended values of .90, the final estimates of the CIPPTTCP scales indicate that all 85 items could be kept and no items was deleted.

As shown in Appendix D, all estimates are significant with Critical Ration (CR) CR>1.96, P-value <0.05, and all the error variance (SE) ≤1.0, indicating no violation of estimates (Al-Shabatat, Abbas & Ismail (2010). Therefore, all 85 items are significantly represented by the 4 variables of the study. After, an estimation of reliability of the subscales was followed through Cronbach's alpha coefficient that yielded .97 for the whole CIPPTTCP questionnaire, which is

a high index of reliability, and .91, .93, .96, and .95 for Content, Input, Process and Product subscales respectively.

Data Analysis

In order to investigate whether there was a significant difference among the respondents' answers to the CIPPTTCP items, all items were closely categorized according to their underlying theme. Since the items under each category are slightly correlated, Pallant (2013) suggests using a single analysis instead of comparing the items separately. Thus, the items that measured different aspects of one theme, or themes closely related to one another, were categorized together and were subjected to the Multivariate Analysis Of Variance (MANOVA) on SPSS. Using MANOVA helps identify if the groups differ significantly in each category and if so in which specific item/s.

Observation Phase

Participant: For the purpose of observations, two courses were chosen using convenience sampling. A total of 28 trainees and 30 different trainers were observed.

Procedure: The researcher sat through two entire courses without any interference to get a clear understanding of the course. While observing, the researcher took notes and filled out the process section of the CIPPTTCP scale.

Journals: The trainees were asked to keep journals during the course. Instructions about how to write the journals, i.e. 'takeaways', were given on the first day of the course. Trainees needed to reflect upon and anonymously write what they had learned, the quality of the material, the trainers' performance, and issues experienced. Then, 50 takeaways were randomly selected, typed and coded using MaxQDA software.

Results

Qualitative Phase Results

After the transcription of interviews and open, axial and then selective coding (Strauss and Corbin, 1997), the frequencies of the codes were 569. Table 3 shows the final tally in detail.

Table 3

Final Tally of Codes According to the CIPP

	Factor	Frequency
1	Context	101
2	Input	130
3	Process	156
4	Output	182
	Total	569

One big concern of stakeholders and a reoccurring theme was considering learner needs and differences:

Excerpt #2:

> *"Since the people taking English classes differ in terms of their interests, background knowledge and abilities, the course must prepare trainees to respond to learners' needs effectively."*

Another theme, which was frequently referred to, was lesson planning:

Excerpt #3:

> *"The graduate of this program must see the importance of showing up to class fully prepared. That is, they need to be on top of what they have taught before, what they are going to cover this particular session, what possible issues they may face and how they will deal with those anticipated problems."*

The analysis and data saturation of trainers' interview transcripts led to the emergence of a number of themes. The themes together with their frequencies are summarized in Table 4.

Table 4

Needs and Expectations from Trainees' Views

Theme	Frequency
I expect to learn about …	
Lesson planning	32
Learner differences	26
Effective teaching	38
Learner types	30
Class management	40
Assessment	31
Creativity	23
Providing feedback	33
Needs analysis	29
Being motivating	31
Being firm	27
Being helpful	37
Teaching language skills and components	49
Enhancing learning	60
Total	486

Quantitative Phase Analysis (CIPPITTCPs Results)

More data on specific objectives set for meeting the trainees' needs were obtained through administrating the questionnaires. Having been closely validated, the questionnaires were given to the participants at all layers of the program. The majority of responses have an inclination towards agreement. The objectives that appeared most in the present dataset are '… train reflective teachers', '… develop trainees' communicative skills.', '… develop trainees'

abilities to maintain appropriate interpersonal relationships with their students.', and '… train motivated teachers.'. These objectives, in alignment with meeting the needs, reached a consensus among the majority of the participants. Overall, all three groups were in agreement regarding the 13 items of this sub-scale.

To investigate if the perspectives of different groups of respondents differed regarding the course objectives, a One-way Between groups Multivariate Analysis of Variance (MANOVA) was used. The dependent variables were the 13 items/dimensions of the CIPPITTCs, which accounted for individual objectives of the course combined with the needs and expectations of the trainers. As mentioned, these items measure the features of an ideal graduate. Preliminary assumption testing was conducted to check for normality, linearity, univariate and multivariate outliers, homogeneity of variance-covariance matrices and multicollinearity. Since some of the assumptions were not fully met, instead of Wilk's Lambda, Pillai's Trace was chosen since it is more robust. It is worth mentioning that Wilk's Lambda and Pillai's Trace are positive statistical values used to ascertain if groups are significantly different or not.

As to the dependent variable items (specific objectives taken as levels in this study), the Multivariate test result appeared significant ($F_{(26, 338)}$, p = .042, Pillai's Trace = .21; Partial eta squared = .01). In order to analyze the results for the dependent variables means conducting several analyses which increases the probability of Type I error, it is suggested by Fidell and Tabachnick (2003) that the original alpha value of .05 be divided by the number of dependent variables. This method is known as a Bonferroni adjustment. Hence the dependent variables were analyzed separately using a Bonferroni adjusted alpha level of .004, the only difference observed was seen in item 7 'Train patient teachers' ($F_{(2, 180)}$, p = .004; Partial eta squared = .06). A closer inspection of the mean scores indicated that the trainers' responses leaned more towards partial agreement (X=2.2, SD = .156) while teachers' and managers' and supervisors' responses leaned towards an agreement, meaning that the latter believed in the attainment of the course objectives and meeting the trainees' needs more than the former group (i.e., trainers). The mean score of the trainers was 2.6 (SD=.09) compared to that of the managers and supervisors (X=2.9; SD = .148).

Overall Objectives

The program was also assessed regarding its overall goals and objectives including 'clarity', and 'measurability'. In total, trainers' and supervisors' responses indicated an agreement. The only item that did not come up with a complete agreed-upon level was the last item upon which all groups partially agreed. A MANOVA was used to investigate if the perceptions of different respondents differed regarding the objectives of the course. The dependent variables were the 5 items/dimensions of CIPPITTCPs which measured the overall objectives of the course. Preliminary assumption testing was conducted and Pillai's Trace was chosen since it is more robust. As to the overall objectives, the Multivariate test was not significant ($F_{(8, 366)}$, p = .103, Pillai's Trace = .071; Partial eta squared = .035), meaning that all participants held similar views regarding the overall objectives of the program which indicated agreement. The last item of the sub-scale must be excluded from the previous statement since all groups only partially agreed with it.

Input Evaluation

Official Document Analysis: The results of the content analyses of the official documents showed a number of themes about the program's input entries, especially the trainers. This list mainly included the following qualifications: 'Pedagogical knowledge', 'Interpersonal skills', 'Supportive', 'Up-to-date knowledge', 'Command of English', 'Passion'. A MANOVA test was employed to compare the CIPPTTCP's respondents' views. This time, the dependent variables were the 10 items/dimensions of CIPPITTCPs scale, which investigated the quality of the trainers' practice. Before the main test, a number of pre-requite tests were run to check the preliminary assumptions of normality, linearity, univariate and multivariate outliers, homogeneity of variance-covariance matrices and multicollinearity. Since some of the assumptions were not fully met, instead of Wilk's Lambda, Pillai's Trace was chosen because of its robustness. The results of the Multivariate test were not significant for the dependent variable (i.e., qualifications and features of trainers) ($F_{(18, 358)}$, p = .071, Pillai's Trace = .145; Partial eta squared = .072), meaning that the respondents were not different in their perceptions and views about the trainers as one of the most fundamental entries to the program.

To investigate if the material and the other facilities are considered as adequate by the program's different beneficiaries, a One-way Between groups MANOVA was used. This time, the dependent variables were the 5 items/dimensions of CIPPITTCPs. The results of the preliminary assumption testing revealed that some of the assumptions were not fully met, therefore, Pillai's Trace was chosen for its robustness. The results of the Multivariate test were significant for the Dependent variable (i.e., material and facilities this time) ($F_{(10, 366)}$, p = .008, Pillai's Trace = .126; Partial eta squared = .06). However, following the procedure explained in 4.3.2.2 regarding the recommendations of Fidell and Tabachnick (2003), a separate analysis of the dependent variables as carried out using a Bonferroni adjusted alpha level of .01 wherein no significant difference was seen.

Process Evaluation

Surveys and Questionnaire Data: As fully described in Methods, trainers were interviewed about what is taught during the program and how, the codes extracted regarding the methodology of the course are shown in Table 5.

Further data about the program process were collected through administering the CIPPTTCPs to trainers, graduated trainees, and branch supervisors. The results show that most responses are clustered around 'agreement' or 'complete agreement'. However, trainers' responses to the 'error correction' and 'needs analysis' items signify a level of 'disagreement' or 'partial agreement'. As to the importance of other categories of the Process component, three adopted strategies of this phase, i.e., 'Assessment', 'Notifications', and 'Self-expressions' were also subjected to further analyses. The results show that the highest percentage was observed for an agreement on the 'Assessment' items.

Table 5

Process Phase: Methodology-related Themes and their Frequencies

Themes	Frequency
Error correction	16
Lesson planning	19
Monitoring	10
Concept Check Questions (CCQs)	11
Instruction Check Questions (ICQs)	8
Teacher talk	7
Student talk	9
Pair and group work	10
Learning styles	6
Needs analysis	13
Importance of personalization	7

However, when exploring any probable difference among the participants' perceptions, the MANOVA test results appeared significant (F (6, 372), p = 0, Pillai's Trace = 4.5; Partial eta squared = .07). The results were separately analyzed for the dependent variable using a Bonferroni adjusted alpha level of .01, and significant differences were observed among respondents' perceptions regarding all three items. For the item 'formative assessment' the mean score for trainers, managers and supervisors, and teachers were found to be 2.35 (SD = .1), 3.0 (SD = .1) and 3.1 (SD = .14) respectively. For the item 'effectiveness of formative assessment of trainees' performance' the mean scores were 2.29 (SD = .12), 3.02 (SD = .07), and 3.0 (SD = .12) respectively. Moreover, for the item 'summative assessment' the results were found to be 2.29 (SD = .12), 2.99 (SD = .74) and 2.88 (SD = .12) for the three respective groups of trainers, managers and supervisors, and teachers.

For 'Notification', ubiquitous agreement in the clarity of the notifications and announcements given during the program were reported by the participants. Likewise, they showed their agreement on the category 'Self-expression'. When the MANOVA tests were run for the two items, no significant differences were observed between the participants (p <.05, F (2, 187) = .48, p = .61; p <.05, F (2, 187) = 2.26, p = .1 for Notification and Self-expression respectively), meaning that participants perceptions did not differ significantly.

Observations: The results showed that while most items were assessed with "agreement", items referring to class management have received more partial agreement compared to other items.

Journals: All journals were transcribed and analyzed for underlying themes using open, axial and selective coding process. Codes were openly created, grouped together and categorized under the 'trainer-, 'session- and 'materials-related' categories. Table 6 reports the themes extracted from the takeaways along with their frequencies.

Table 6

Themes Extracted from Takeaways and their Frequencies

Category					
Trainers		Sessions		Material	
Code	*f*	*Code*	*f*	*Code*	*F*
Friendly	97	Positive atmosphere	97	Clear	90
Supportive	92	Well-planned	99	Helpful	98
Prepared	80	Informative	100		
Energetic	96	Practical	98		
Confusing	5	Interesting	90		
Confident	90	Boring	2		
Irritable	3	Productive	94		
Caring	86	Fun	92		
Organized	98				
Creative	91				

Note: f=frequency

Product Evaluation

Impact: A combination of the data obtained through the thematic analyses of the interviews with the trainees, the takeaways, and the available literature was used in this stage. The outcome formed the basis for developing the required evaluation items. Another MANOVA test was run on the present 8 items of the Impact sub-component. After preliminary assumption testing, Pillai's Trace was chosen due to its robustness. The results were significant for the variable of Impact ($F (16, 360)$, $p = .000$, Pillai's Trace $= .137$; Partial eta squared $= .01$). Analyzing the results for the dependent variable separately and using a Bonferroni adjusted alpha level of .004, differences were observed for item Prod 5 'Trainees have a good understanding of management (class and time)' with $F (2, 186)$, $p = 0$; Partial eta squared $= .082$. A comparison of the mean scores of different groups shows that on average trainers only partly agree with this statement ($X = 2.21$, $SD = .148$). Almost the same results were observed for managers and supervisors ($X = 2.4$, $SD = .086$). However, teachers mean score was higher than those of the others ($X = 2.98$, $SD = .14$). Regarding item Prod 7 'the program increases trainees' general knowledge in English' ($F (2, 186)$, $p = .002$; Partial eta squared $= .066$), mean scores showed that the trainers, on average, disagreed with this statement ($X = 1.48$, $SD = .169$) while managers/supervisors and teachers partly agreed ($X= 2.08$, $SD =.098$, and $X =2.28$, $SD = .158$), respectively. With respect to item Prod 12 'Trainees have learned about curricular planning and instruction' ($F (2, 186)$, $p = 0$; Partial eta squared $= .102$), a closer observation of the different groups showed that while the teachers only partly agreed with this item ($X = 2.19$, $SD = .133$), the managers and supervisors ($X = 2.86$, $SD = .077$) and the trainers ($X = 2.88$, $SD = .124$) have a more positive opinion about the trainees achieved skills in this regard.

Effectiveness: Considering participants' different perspectives on the effectiveness, the results of a MANOVA test appeared significant ($F (20, 344)$, $p = .02$, Pillai's Trace $= .183$; Partial eta squared $= .09$). The results for the dependent variable were analyzed separately and using a Bonferroni adjusted alpha level of .005, no significant difference was observed.

Sustainability: The results of MANOVA test for differences appeared non-significant (F (8, 362), p = .09, Pillai's Trace = .07; Partial eta squared = .04) indicating that the participants perceptions regarding the sustainability of the outcomes do not significantly differ.

Transportability: Respondents were asked about the program's applicability in a new context and the highest percentage of responses were '*agree*' or '*strongly agree*'. The perceptions of different participants regarding the program's transportability compared using a MANOVA test. The multivariate result was significant for transportability sub-component (F (8, 362), p = .033, Pillai's Trace = .09; Partial eta squared = .045). However, a closer analysis of the dependent variables separately and with a Bonferroni adjusted alpha level of .01 revealed no significant difference, meaning that they all believed in the program's transportability and applicability in a new context.

Discussion

To provide evidence required for informed decision to serve program quality and improvement, many authors (e.g.Greene, 1988; Ryan & Cousins, 2009) argue evaluation should expand to include meaningful explanation on both policy and practice of the program. Given this, the present evaluation undertaken in the context of Safir Language Academy, Iran, as one of the most populated language schools of the country, aimed to investigate the teacher-training program delivered by this specific Academy. The present evaluation approach was carried out through the Stufflebeam's CIPP framework and aimed to examine the program's objectives, effectiveness, impact, sustainability and transportability. What follows is a discussion of the findings in relation to these areas and the main research questions.

Context

According to the results of the qualitative and quantitative analyses of the obtained data, the present course reveals to have clearly set objective. The goals of the program are in line with those of CLT and some organizational values taken from the literature' (Brown & Lee, 1994; Harmer, 2008). The agreement observed among all stakeholders in this regard corroborates this conclusion. The results showed no significant difference between the opinions of teachers and trainers regarding the objectives except for item C 7 '… train patient teachers' showing that the trainers do not necessarily agree with this objective but the supervisors and teachers do. This discrepancy could mean that while training patient teachers is not necessarily the goal of the program, it is indirectly communicated to the trainees and hence has become highlighted as a goal in their perceptions.

As to the objectives of the program, the analysis of documents and interviews showed they are clear, measurable and constructive. This assumption was also confirmed by the responses given to the related items in the CIPPTTCP scale. Further analysis of the results showed no significant difference among the perceptions of different respondents. Furthermore, results showed the insufficiency of the program's duration. This was confirmed by CIPPTTCP's results. All respondents majorly exhibit partial agreement or even disagreement. This could mean that the program's policymakers might need to lengthen the course. This issue seems to have surfaced in similar studies of teacher education or training (Masoumpanah et al., 2017).

Thus, sufficient evidence exists to convince policymakers towards allocating more practice time to their preparation programs.

Input

The results showed that the standards defined for a qualified trainer are clearly set. These standards are consistently maintained by the trainers. All groups of respondents agreed that the trainers are knowledgeable, active, supportive, and up-to-date, with no significant difference observed among respondents. Qualitative analysis of the trainees' journals showed that the trainers' performance is in line with the intended trainer-portfolio explained by the official documents. Furthermore, the response reports showed that the material used is of good quality, sufficient and effective. No significant difference was found, in this regard, among the opinions of different groups.

Process

A close analysis of the data obtained about the program process, clearly indicated that the intended strategies of the program are put into practice as no significant difference was shown among any of the groups' opinions. More corroborating evidence came from the content analysis of the trainees' journals. Further support was provided through the observation data. The results showed a consistency in trainers' teaching styles and their adherence to the standards of the program. Another noteworthy result is that the program's intention to train reflective teachers is clearly put into practice as it was repeatedly evidenced by the trainees' journaling data. In addition, class management, time and crisis, as another intended strategy of the program was reportedly instructed. The results of the MANOVA test showed a significant difference between the perceptions of the trainers and those of the teachers and the supervisors for "crisis management". While the latter had a more affirming outlook towards instruction of this strategy, the former had some doubts regarding its sufficient instruction. This was rather predictable since in the interviews with the trainers, class management was considered as one of the areas needing adjustments. The trainers stated that shortage of time, simulation of a real class, not practicing with real students, all seem to prevent the chance of comprehending the critical situations of real classes. Thus, it was suggested that the trainees have observations of more experienced teachers' classes. This issue was reported to influence error treatment, too. The analysis of this item showed no significant difference among different participants' opinions; however, the qualitative analysis of trainers' interviews clarifies that if trainees were to work with actual students, they would have a better understanding of this skill. Assessment system of the course was another important area explored. The relevant findings for this item showed significant differences between the views of trainers and teachers and supervisors. Trainers' views were less positive perhaps because they expect a more systematic assessment system than others. Regarding the program's announcements, it appears to inform the trainees successfully. All groups agreed upon the item explaining this strategy and no significant difference was observed among them.

Product

Since the purpose of this component of evaluation is ascertaining if the program has satisfied the needs of its target audience (Zhang et al., 2011), the responses given by teachers and supervisors were inspected closely. The results showed that the beneficiaries are generally satisfied with the program and their expectations have been met. However, regarding the 'Management' item, they were not in a complete agreement, and trainers were skeptical of trainees' grasp of class management skills. As to the program's '*impact*' on the trainees' English command, a significant difference was observed among the respondents. While trainers disagreed that there has been any change in this matter, the trainees partly agreed with this item. This can mean that without having intended to, the program has caused an improvement in the trainees' command of English and inspired some trainees to upgrade their knowledge.

Regarding the program's impact on curricular planning, the perspectives differed. The results showed that trainers only *partially agreed* with this effect, while teachers and supervisors on average agreed. It may be because the fixed curriculum does not allow opportunities for this skill being taught. The findings regarding '*effectiveness*' showed that the majority of the participants believed in the significance, quality and acceptability of the program's outcomes. The results of MANOVA showed no significant difference among respondents. Interviews with trainers, trainees' journals and managers' scale responses reflected a good level of satisfaction with the outcomes of the program. Regarding the effect on the trainees' teaching competency, the participants reported a good level of agreement, and no significant difference was found among them. With respect to the program's alignment with the ELT principles, the findings indicated the respondents' *agreement* on this item and no significant difference was observed among different groups. This is an important achievement for the program since the official documents and the stakeholders have placed a vital importance on the strategies being in accordance with ELT standards. Relevantly, the program has successfully managed to increase trainees' confidence to teach independently in class. Personalization of methods, reflection and multiple opportunities to practice teaching are among the reasons listed by trainers and trainees as important contributors to this outcome. Opportunities to practice teaching seemed to create a platform for trainees to receive constructive feedback and get more comfortable speaking in front of people. As a result, many trainers have suggested the number of Teaching Practice sessions be increased. Many of the interviewees reported that in many cases the details focused on during the course (for instance grouping techniques, or designing kinesthetic activities) tend to over-shadow the more fundamental stages of teaching (such as error correction, effective presentation of the lesson). Interestingly, responses of the teachers to this item showed a noticeable level of *partial agreement*. This could mean that the course needs to place a more discernible emphasis on such factors so that trainees can prioritize fundamentals over subtleties.

One of the most important fundamentals of teaching is 'aim achievement' (Brown, 2000), which means designing purposeful tasks and activities. In this study, this fundament was investigated through item Prod 17. The comparison of the results did not yield a tangible significance difference among the responses; however, trainer's interviews showed a concern for the trainees' grasp of this concept. Interestingly, a noticeable percentage of *partial*

agreement was reported by supervisors. This could mean that issue of 'aim achievement' persist when trainees start teaching in branches.

Overall, the present findings provide evidence in support of the effective administration of the program work plan and consensus was reported by different groups. Such a unanimous agreement indicates that the program has been generally successful in achieving its goals through its intended and executed strategies.

Sustainability

Most respondents, showed agreement, and no significant difference was seen in their viewpoints. Indeed, most of them assumed what they acquired during the program could prevail over time. One trainer suggested that the TRD and educational branches work in close conjunction. In this regard, report cards were created after each trainee's final teaching demonstration and sent to their education branches. Furthermore, 'a long-term effect' was explicitly reported by the participants, with no significant differences among their viewpoints. It can be argued that program supervisors and managers play an important role in the sustainability of the standards taught in such programs (Zhang et al., 2011). An important regulation of Safir Language Academy is its frequent observations of teachers by the supervisors of each branch. Each observation is followed by a negotiation session during which feedback is provided on the teacher's performance. Another essential feature of sustainability of the program is the atmosphere of branches. If the values taught in the training programs are highlighted by the existing teachers, the same view can be transferred to the new teachers. This can be corroborated by Farber's Farber (1991) view stating that trainees' enthusiasm and sense of commitment is likely to wither when they are surrounded by negative, burned-out teachers. Similar statements are made by Hoy and Woolfolk (1993) in their discussion of teachers' sense of efficacy and organizational health. Hence, it can be concluded that in order for the effects of the program to sustain, supervisors need to maintain a positive atmosphere in the educational branches. The present data revealed the existence of a post-training course, since teachers have educational workshops regularly. The responses, in this regard, mostly reflect 'agreement' with no significant difference observed.

Transportability

Consensus was observed among the participants, with no significant differences. This might indicate that other contexts could benefit from such courses. However, the standards and techniques of this language school may not be as effective in a public-school context due to a lack of facilities, curricular restrictions, or time constraints. Overall, one possible outcome from the synthesis of all the present findings is evidence for a less positive assessment made by the program trainers than other participants. This, Tschannen-Moran and Hoy (2007) acknowledged, is not necessarily due to the underachievement of the program since they may simply have higher expectations. This assumption is only made stronger by the positive assessments made by the teachers and supervisors. The same results have been reported by Karatas and Fer (2009) as they suggested that such opinions might be due to their higher expectations or deeper grasp of the shortcomings and strengths. Yet, recommendations like this make a case for a consideration of the points raised by trainers.

Conclusion

Evaluation is an inseparable part of all programs since it provides information about the programs' merits and demerits (Zhang et al., 2011). Based on evaluation evidence, informed decision can be made towards program enhancement (Ryan & Cousins, 2009). Given the private sector's role in ELT in Iran, it is important to evaluate their performance. Moreover, since teachers' role in student learning is of an undeniable importance, extra care must be given to their training process (Atai & Mazlum, 2013; Foroozandeh et al., 2007). Hence, the CIPP model formed the framework for the evaluation of one of the largest Iranian language schools, Safir Language Academy. The framework entails careful evaluation of the programs' objectives, the actions planned, the planned strategies in action, and the outcome of the program. To gather evidence, the study proceeded based on a series of in-depth interviews with the participating trainers, trainees or other stakeholders. The outcome of this qualitative phase consequently led to the development of tools for the quantitative phase, which aimed to probe more extensively into the program's stakeholders. As Stufflebeam and Shinkfield (2007) argue for the triangulation of the information, the data gathered in this study were also triangulated through close interviews, observations, surveys, journals, and content analysis of the official documents. The overall evaluation phase yielded the following results about the program. Firstly, the results revealed that the program does have clear objectives consistent with the needs and expectations of the trainees. Secondly, the analysis of input in terms of its methods, plans and procedures and other resources were examined showed a positive perspective held by all the beneficiaries. In addition, the 'hows' of implementing the plans were assessed. Although some minor discrepancies were observed among the views, in general, the strategies proved to be effectively put into practice. Finally, the analysis of the program's impact, sustainability, transportability and effectiveness showed that the outcomes were perceived to be satisfactory, of an acceptable quality, long lasting and applicable elsewhere. However, degrees of minor differences were also observed. The evaluation showed consistent, energetic, helpful and well-planned performance from the trainers leading to an indirect but significant indication of 'how the ideal teacher is supposed to act'.

This evaluation suggests that more time needs to be allocated to the program since the intensity of the input may overwhelm the trainees and leave little time for practicing the input. Additionally, it was observed that 'simulated classroom', i.e., not having real students, might result in an insufficient grasp of the class management, time management and error correction. This study lasted for almost a year; however, a longitudinal study of the programs' outcomes is suggested. In this way, more time is provided for the analysis of participants' values, attitudes and views and for comparing them with each other. Moreover, a more detailed analysis of the curriculum designed and the materials, i.e., books, worksheets, etc., is suggested for further studies.

ORCID

https://orcid.org/0009-0007-1578-1478

https://orcid.org/0000-0001-6949-4916

https://orcid.org/0000-0002-5713-7062

Rights and Permissions

Open Access

References

Atai, M. R., & Mazlum, F. (2013). English language teaching curriculum in Iran: Planning and practice. *The Curriculum Journal*, *24*(3), 389-411. https://doi.org/10.1080/09585176.2012.744327

Barry, D., Pendergast, D., & Main, K. (2020). Teacher perspectives on the use of the Australian professional standards for teachers as part of their evaluation Process. *Australian Journal of Teacher Education*, *45(8)*, 1-22. https://doi.org/10.14221/ajte.2020v45n8.1

Berg, D. A. G., Skaalvik, E. M., Asil, M., Hill, M. F., Uthus, M., Tangen, T. N., & Smith, J. K. (2023). Teacher self-efficacy and reasons for choosing initial teacher education programmes in Norway and New Zealand. *Teaching and Teacher Education*, *125*, 104041. https://doi.org/10.1016/j.tate.2023.104041.

Brown, H. D. (2000). Principles of language learning and teaching. Longman.

Brown, H. D., & Lee, H. (1994). *Teaching by principles: An interactive approach to language pedagogy* (Vol. 1). Prentice Hall Regents Englewood Cliffs, NJ.

Celce-Murcia, M., & McIntosh, L. (1991). *Teaching English as a second or foreign language* (3rd Edition). Heinle & Heinle Publisher.

Chambless, K. S. (2012). Teachers' oral proficiency in the target language: Research on its role in language teaching and learning. *Foreign Language Annals*, *45*(1), 141-162. https://doi.org/10.1111/j.1944-9720.2012.01183.x

Clark, S., & Newberry, M. (2019). Are we building preservice teacher self-efficacy? A large-scale study examining teacher education experiences. *Asia-Pacific Journal of Teacher Education*, *47*(1), 32-47. http://dx.doi.org/10.1080/1359866X.2018.1497772

Crandall, J. (2000). Language Teacher Education. *Annual Review of Applied Linguistics*, *20*, 34-55. http://dx.doi.org/10.1017/S0267190500200032

Daniel, J. (2011). *Sampling essentials: Practical guidelines for making sampling choices*. Sage.

Darling-Hammond, L., & Bransford, J. (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. John Wiley & Sons.

Darussalam, G. (2010). Program evaluation in higher education. *International Journal of Research & Review*, *5*(2), 56–65. https://journal.translationstudies.ir/ts/article/view/800

Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, *99*(3), 597-610. https://psycnet.apa.org/doi/10.1037/0022-0663.99.3.597

Dursun, H., Agirdag, O., & Claes, E. (2023). Unpacking preservice teachers' beliefs about students' home languages: what matters in initial teacher education? *Journal of Multilingual and Multicultural Development*, 1-16. https://doi.org/10.1080/01434632.2023.2173209

Ellis, V., & Childs, A. (2019). Innovation in teacher education: collective creativity in the development of a teacher education internship. *Teaching and Teacher Education*, *77*, 277-286. https://doi.org/https://doi.org/10.1016/j.tate.2018.10.020

Farber, B. A. (1991). *Crisis in education: Stress and burnout in the American teacher*. Jossey-Bass.

Farrell, T. S. (2011). Exploring the professional role identities of experienced ESL teachers through reflective practice. *System*, *39*(1), 54-62. http://dx.doi.org/10.1016/j.system.2011.01.012

Farrell, T. S. (2012). Novice-service language teacher development: Bridging the gap between preservice and in-service education and development. *TESOL Quarterly*, *46*(3), 435-449. https://doi.org/10.1002/tesq.36

Fidell, L. S., & Tabachnick, B. G. (2003). Preparatory data analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology,* Vol. 2, pp. 115–141). John Wiley & Sons, Inc. https://doi.org/10.1002/0471264385.wei0205

Foroozandeh, E., Riazi, A., & Sadighi, F. (2007). TEFL program evaluation at master's level in Iran. *Teaching English Language*, *2*(2), 71-100. https://doi.org/ 10.22132/TEL.2009.122663

Freeman, D., & Johnson, K. E. (1998). Reconceptualizing the knowledge-base of language teacher education. *TESOL Quarterly*, *32*(3), 397-417. https://doi.org/10.2307/3588114

Frye, A. W., & Hemmer, P. A. (2012). Program evaluation models and related theories: AMEE Guide No. 67. *Medical Teacher*, *34*(5), 288-299. https://.doi.org/10.3109/0142159x.2012.668637

Ghasemi, N., Najafi, E., Lotfi, F. H., & Sobhani, F. M. (2020). Assessing the performance of organizations with the hierarchical structure using data envelopment analysis: An efficiency analysis of Farhangian University. *Measurement*, *156*, 107609. https://doi:10.1016/j.measurement.2020.107609

Glaser, B. G., & Holton, J. (2004). Remodeling Grounded Theory. *Forum Qualitative Sozialforschung Forum: Qualitative Social Research*, 5(2). 1-22. https://doi.org/10.17169/fqs-5.2.607

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Aldine Transaction.

Greene, J. C. (1988). Communication of results and utilization in participatory program evaluation. *Evaluation and Program Planning*, *11*(4), 341-351. https://doi.org/10.1016/0149-7189(88)90047-X

Grossman, P. (2021). *Teaching core practices in teacher education*. Harvard Education Press.

Harmer, J. (2008). How to teach English. *ELT Journal*, *62*(3), 313-316. https://doi.org/10.1093/elt/ccn029

Harris, & Sass. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7-8), 798-812. https://doi.org/r/eee/pubeco/v95y2011i7-8p798-812.html

Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *The Elementary School Journal*, *93*(4), 355-372. https://psycnet.apa.org/doi/10.1086/461729

Iranmehr, A., & Davari, H. (2018). English language education in Iran: A site of struggle between globalized and localized versions of English. *Iranian Journal of Comparative Education*, *1*(2), 94-109. https://doi.org/10.22034/ijce.2018.87725

Karatas, H., & Fer, S. (2009). Evaluation of English curriculum at Yildiz Technical University using CIPP model. *Egitim ve Bilim*, *34*(153), 47. https://www.researchgate.net/publication/298713754

Kaufman, R., & Keller, J. M. (1994). Levels of evaluation: Beyond Kirkpatrick. *Human Resource Development Quarterly*, *5*(4), 371-380. https://www.researchgate.net/publication/298713754

Kirkpatrick, D. (1996). Revisiting Kirkpatrick's four-level model. *Training & Development*, *50*(1), 54-57.

Kline, R. (1998). *Principles and practice of SEM*. The Guilford Press.

Koenig, HG, Kvale, JN, & Ferrel, C.(1988). Religion and well-being in later life. *The Gerontologist*, *28*, 18-28. https://psycnet.apa.org/doi/10.1093/geront/28.1.18

Korthagen, F. A., Kessels, J., Koster, B., Lagerwerf, B., & Wubbels, T. (2001). *Linking practice and theory: The pedagogy of realistic teacher education*. Routledge.

Levine, T. H., Mitoma, G. T., Anagnostopoulos, D. M., & Roselle, R. (2023). Exploring the nature, facilitators, and challenges of program coherence in a case of teacher education program redesign using Core practices. *Journal of Teacher Education*, *74*(1), 69-84. http://doi.org/10.1177/00224871221108645

Loewenberg Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389-407. https://doi.org/10.1177/0022487108324554

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First-and higher order factor models and their invariance across groups. *Psychological Bulletin*, *97*(3), 562-582. https://psycnet.apa.org/doi/10.1037/0033-2909.97.3.562

Masoumpanah, Z., Tahririan, M. H., Alibabaee, A., & Afzali, K. (2017). Evaluation of the Undergraduate TEFL Program at Farhangian University: Merits and Demerits. *Iranian Journal of Applied Linguistics*, *20*(2), 157-193. http://ijal.khu.ac.ir/article-1-2826-en.html

McNamara, C. (2008). *Basic guide to program evaluation*. Free Management Library.

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

Pallant, J. (2013). *SPSS survival manual*. McGraw-Hill Education.

Peacock, M. (2009). The evaluation of foreign-language-teacher education programmes. *Language Teaching Research*, *13*(3), 259-278. https://doi.org/10.1177/1362168809104698

Ping, C., Schellings, G., & Beijaard, D. (2018). Teacher educators' professional learning: A literature review. *Teaching and Teacher Education*, *75*, 93-104. https://doi.org/10.1016/j.tate.2018.06.003

Ryan, K., & Cousins, J. B. (2009). *The sage international handbook of educational evaluation*. Sage.

Safir-Language-Academy. (2023). *Safir's English language teacher training course*. https://gosafir.com/fa/

Safir-Language-Academy. (2023). *Units of Safir Language Academy*. https://gosafir.com/fa/safir-branches/

Schacter, J., & Thum, Y. M. (2004). Paying for high-and low-quality teaching. *Economics of Education Review*, *23*(4), 411-430. http://dx.doi.org/10.1016/j.econedurev.2003.08.002

Strauss, A., & Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Stufflebeam, D. (2004). *The 21st-century CIPP model: Origins, development, and use*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984157.N16

Stufflebeam, & Shinkfield. (2007). Evaluation theory, models, and applications. Jossey-Bass.

Tschannen-Moran, M., & Hoy, A. W. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education*, *23*(6), 944-956. https://doi.org/10.1016/j.tate.2006.05.003

Von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, *53*, 31-45. http://www.sciencedirect.com/science/article/pii/S0272775716302503

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines*. White Plains, NY: Longman.

Young Lee, S., Shin, J.-S., & Lee, S.-H. (2019). How to execute Context, Input, Process, and Product evaluation model in medical health education. *Journal of Educational Evaluation for Health Professions*, *16, 40-0.* https://doi.org/10.3352%2Fjeehp.2019.16.40

Zhang, G., Zeller, N., Griffith, R., Metcalf, D., Williams, J., Shea, C., & Misulis, K. (2011). Using the context, input, process, and product evaluation model (CIPP) as a comprehensive framework to guide the planning, implementation, and assessment of service-learning programs. *Journal of Higher Education Outreach and Engagement*, *15*(4), 57-84.

Appendices

Appendix A

Interview Guide for Trainers

1. What are we training to achieve in our trainees?

2. Who do you consider to be a good teacher?

3. What are the features of a good teacher trainer?

4. Are the planned strategies being implemented? Or how are the planned strategies being implemented?

5. Who do you consider as a successful graduate of the course?

6. Do you think the points taught during the course as important to the teachers after the course ends?

7. How can we make the points taught in the course more sustainable?

Appendix B

Interview Guides for Trainees

1. What do you consider the qualifications of a good teacher to be?

2. What do expect to learn in the course?

3. What do you need the course to prepare you for?

Appendix C

The CIPPTTCP Scale

Dear colleague,

The purpose of this study is to evaluate the quality and effectiveness of XX Institute's TTC program, as a whole, and from the perspective of its intended goals, product, process, success and impact. Your time and kind contribution to this study are greatly appreciated.

Section I: Evaluation Form

Please tick the one choice below that you think best describes the program.

The **context** Component of the Program															
	Respondents														
	Trainers					**Trainees**					**Evaluator**				
Statement	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4
Specific Objectives: The program objective is to ….															
1. train reflective teachers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
2. develop trainees' communicative skills.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
3. train supportive/encouraging teachers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
4. develop trainees' problem-solving skills.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
5. develop trainees' teaching autonomy.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
6. train creative teachers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
7. train patient teachers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
8. train self-conscious teachers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
9. train flexible teachers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
10. develop trainees' abilities to maintain appropriate interpersonal relationships with their students.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
11. develop trainees' teaching ethics.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
12. train motivated teachers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
26. develop trainees' skills for active teaching.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Overall Objectives and Goals															
30. The program has clear objectives.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
31. The program objectives are measurable.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
32. The program objectives meet the trainees' needs.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4

33. The program is adequate for the improvement of trainees' professional development.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
34. The duration of the program is sufficient for achieving its goals.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4

The **Input** Component of the Program															
Statements **In this program………**	Respondents														
	Trainers					**Trainees**					**Evaluator**				
	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4
3. Trainers are competent enough to train teaching methodology and strategies to trainees.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
4. Trainers have interpersonal skills.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
5. Trainers are encouraging/supportive.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
6. Trainers' knowledge is up-to-date.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
7. Trainers have a good command of English.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
8. Trainers have passion for training the trainees.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
9. Organizational rules are consistently followed/observed by trainers.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
10. Trainers have active presence.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
11. Trainers deliver effective training.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
12. Trainees have the required knowledge base.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
13. Trainees have the positive drive to participate in the program.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
18. The program applies ELT jargon into its training.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
19. The program delivers up-to-date teaching theories to trainees.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
20. Trainers provide trainees with sufficient handouts and written materials.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
21. Trainers deliver their training using technology equipment (e.g., projector, slides, etc.).	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
28. The program classwork help trainees learn easily.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Overall Input Evaluation															

The **Process** Component of the Program															
	Respondents														
	Trainers					**Trainees**					**Evaluator**				
Statement **In this program……………**	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4
8. 'Error correction' techniques are instructed.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
9. 'Lesson planning' is taught.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
10. 'Teacher monitoring' is taught.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
11. Trainers teach about Making and asking Concept Check Questions (CCQs).	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
12. Trainers teach about Making and asking Instruction Check Questions (ICQs).	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
13. Trainers teach about the amount and importance of 'Teacher talk'.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
14. Trainers teach about the amount and importance of 'Student talk'.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
15. Trainees' pair and group works rather than individual work alone are included.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
16. Trainers teach about various 'students' pair and 'group' works.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
17. 'Teacher time management' is instructed.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
18. 'Teacher crisis management' is instructed.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
19. 'Teacher discipline' is instructed.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
21. Learners' different learning styles' are described and taught.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
22. 'Effective teacher observation' is instructed.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
23. 'Reflective teaching' is instructed.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
24. 'Need analysis strategy' is instructed.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
25. The importance of 'personalization of learning' is emphasized.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
26. 'Teacher autonomy' is fostered.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
27. 'Organizational culture/values' are taught.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
28. 'Problem solving skills' are taught.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4

29. Techniques for establishing a 'positive learning environment' are taught.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
30. Techniques for 'student assessment and evaluation' are provided.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
31. Notifications and announcements are clearly given.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
32. Opportunities for trainees' self-expression are provided.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
34. Trainees have active participation.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
35. Trainees are assessed formatively.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
36. The assessment affects trainees' performance positively.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
37. Trainees are assessed summatively.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4

The **Product** component of the program.

Statement	Respondents														
	Trainers					Trainees					Evaluator				
	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4	Strongly disagree 0	Disagree 1	Partly agree 2	Agree 3	Strongly agree 4
Impact: Reach to the target audience															
5. Trainees have a good understanding of management (class and time).	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
6. The program motivates trainees to update their knowledge.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
7. The program increases trainees' general knowledge in English.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
8. The program increases trainees' content knowledge in English.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
9. Trainees become aware of learner related factors.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
10. Trainees become aware of teacher related factors contributive to student learning.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
11. Trainees have acquired techniques for effective teaching of language skills and components.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
12. Trainees have learned about curricular planning and instruction.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Effectiveness: Significance and quality of the outcomes															
14. Trainees' teaching competency has satisfactorily improved.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4

15. Trainees' teaching performance has satisfactorily improved.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
16. Trainees can satisfactorily prioritize fundaments of teaching over the details of teaching.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
17. Trainees can prioritize 'goal achievement' over the details of teaching.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
18. Trainees' abilities in teaching language skills and components have satisfactorily improved.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
19. Trainees have satisfactorily learned adhering to ELT discipline.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
20. Trainees have satisfactorily learned when and how to provide learners with feedback.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
21. Trainees have grasped the importance of student learning.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
22. Trainees' confidence to work independently has improved.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
28. The program work plan has been effectively administered.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Sustainability: how the results are continued over time															
29. The program contribution to trainees' professional development continues over time.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
34. The points learned during the program have a long-term effect.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
35. The points taught during the program are of great value and importance to trainees.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
36. There is a post-training program for novice teachers to assess their learning.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Transportability: program's adaptability in relevant context.															
37. The program can yield similar results in a new context.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
38. The program content can be beneficial in new context.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
39. The adopted methods can be applied in a new context.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
40. The program work plan can be applied in new context.	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4

Section II. Personal Information Collected

- *Gender:_____*
- *Age:*
- *University Major(s):*
- *Professional Degree (s) Achieved (BA, MA. PhD.):*
- *Years of teaching:_____*

- *Position (job title) at XX Institute:_____*
- *Please write your email below! We may need it! Thank You!*

Appendix D

Parameter Estimates of the Standardized Factor Loadings, Standard Error (SE), and Critical Ratio (CR) for the Measurement Model (CIPP)

			Estimate	S.E.	C.R.	P	Label
Q1C	<---	Context	1.000				
Q2C	<---	Context	.770	.108	7.151	***	par_1
Q3C	<---	Context	.925	.110	8.447	***	par_2
Q4C	<---	Context	1.138	.125	9.095	***	par_3
Q5C	<---	Context	1.310	.137	9.581	***	par_4
Q6C	<---	Context	1.057	.129	8.195	***	par_5
Q7C	<---	Context	1.113	.126	8.805	***	par_6
Q8C	<---	Context	1.097	.119	9.214	***	par_7
Q9C	<---	Context	1.044	.119	8.788	***	par_8
Q10C	<---	Context	.770	.098	7.845	***	par_9
Q11C	<---	Context	1.080	.126	8.562	***	par_10
Q12C	<---	Context	.854	.103	8.276	***	par_11
Q26C	<---	Context	.734	.113	6.469	***	par_12
Q30C	<---	Context	.642	.094	6.861	***	par_13
Q31C	<---	Context	.620	.104	5.985	***	par_14
Q32C	<---	Context	.886	.100	8.901	***	par_15
Q33C	<---	Context	.950	.120	7.892	***	par_16
Q34C	<---	Context	.662	.124	5.339	***	par_17
Q3I	<---	Input	1.000				
Q4I	<---	Input	.931	.073	12.732	***	par_18
Q5I	<---	Input	.794	.066	12.093	***	par_19
Q6I	<---	Input	.962	.079	12.231	***	par_20
Q7I	<---	Input	.712	.068	10.485	***	par_21
Q8I	<---	Input	.967	.077	12.514	***	par_22
Q9I	<---	Input	1.000	.074	13.580	***	par_23
Q10I	<---	Input	1.016	.076	13.418	***	par_24
Q11I	<---	Input	.934	.064	14.595	***	par_25
Q12I	<---	Input	.969	.090	10.810	***	par_26
Q13I	<---	Input	.777	.076	10.280	***	par_27
Q18I	<---	Input	.556	.082	6.735	***	par_28
Q19I	<---	Input	.639	.090	7.113	***	par_29
Q20I	<---	Input	.525	.079	6.676	***	par_30
Q21I	<---	Input	.710	.073	9.752	***	par_31
Q28I	<---	Input	.656	.074	8.866	***	par_32
Q8PROC	<---	Process	1.000				
Q9PROC	<---	Process	.921	.076	12.084	***	par_33
Q10PROC	<---	Process	.807	.068	11.821	***	par_34
Q11PROC	<---	Process	.820	.073	11.272	***	par_35
Q12PROC	<---	Process	.809	.071	11.328	***	par_36
Q13PROC	<---	Process	.924	.075	12.370	***	par_37
Q14PROC	<---	Process	.717	.071	10.166	***	par_38
Q15PROC	<---	Process	.508	.061	8.400	***	par_39
Q16PROC	<---	Process	.526	.059	8.961	***	par_40
Q17PROC	<---	Process	.858	.070	12.294	***	par_41
Q18PROC	<---	Process	1.042	.086	12.065	***	par_42
Q19PROC	<---	Process	.980	.083	11.786	***	par_43
Q21PROC	<---	Process	.863	.075	11.486	***	par_44
Q22PROC	<---	Process	1.045	.080	13.122	***	par_45
Q24PROC	<---	Process	.809	.077	10.530	***	par_46

Maryam Khaksar, Gholam Reza Kiany, Parvaneh ShayesteFar

			Estimate	S.E.	C.R.	P	Label
Q25PROC	<---	Process	.913	.075	12.195	***	par_47
Q27PROC	<---	Process	.907	.075	12.008	***	par_48
Q28PROC	<---	Process	.798	.078	10.225	***	par_49
Q29PROC	<---	Process	.769	.074	10.360	***	par_50
Q30PROC	<---	Process	.943	.086	10.965	***	par_51
Q31PROC	<---	Process	.629	.073	8.579	***	par_52
Q32PROC	<---	Process	.810	.076	10.692	***	par_53
Q35PROC	<---	Process	.776	.077	10.051	***	par_54
Q36PROC	<---	Process	.689	.069	10.045	***	par_55
Q37PROC	<---	Process	.639	.068	9.328	***	par_56
Q5PROD	<---	Product	1.000				
Q6PROD	<---	Product	.901	.102	8.806	***	par_57
Q7PROD	<---	Product	.872	.127	6.841	***	par_58
Q8PROD	<---	Product	.843	.105	8.012	***	par_59
Q9PROD	<---	Product	.887	.097	9.166	***	par_60
Q10PROD	<---	Product	.790	.088	8.945	***	par_61
Q11PROD	<---	Product	.806	.087	9.261	***	par_62
Q12PROD	<---	Product	.707	.102	6.909	***	par_63
Q14PROD	<---	Product	.843	.089	9.521	***	par_64
Q15PROD	<---	Product	.852	.094	9.110	***	par_65
Q16PROD	<---	Product	1.013	.106	9.533	***	par_66
Q17PROD	<---	Product	1.156	.114	10.122	***	par_67
Q18PROD	<---	Product	.938	.092	10.182	***	par_68
Q19PROD	<---	Product	.914	.097	9.449	***	par_69
Q20PROD	<---	Product	.934	.098	9.500	***	par_70
Q21PROD	<---	Product	.788	.101	7.801	***	par_71
Q22PROD	<---	Product	.787	.094	8.396	***	par_72
Q28PROD	<---	Product	.714	.079	9.060	***	par_73
Q29PROD	<---	Product	.857	.106	8.095	***	par_74
Q34PROD	<---	Product	.928	.104	8.932	***	par_75
Q35PROD	<---	Product	.733	.084	8.702	***	par_76
Q36PROD	<---	Product	.846	.132	6.405	***	par_77
Q37PROD	<---	Product	.838	.087	9.664	***	par_78
Q38PROD	<---	Product	.750	.081	9.253	***	par_79
Q39PROD	<---	Product	.696	.080	8.723	***	par_80
Q40PROD	<---	Product	.754	.081	9.339	***	par_81