LANGUAGE PROGRAM EVALUATION: A REFLECTION ON SOME EVALUATION DESIGNS

Muhammad Salim Tufail [1]*, Melor Md Yunus [2]

[1] Language Centre, National Defence University of Malaysia, Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia

[2] Faculty of Education, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

*For correspondence; Tel. + (60) 0390513004, Email: salim@upnm.edu.my

ABSTRACT: *One of the key areas to consider when one decides to conduct a language program evaluation is to select an evaluation design. Based on the work of leading experts, there are generally three broad categories of evaluation designs. The first one is called the positivist evaluation design, where phenomena are measured and supported by objective evidence. Being closely linked to the "quantitative dimension" it is usually summative in nature or product oriented. It is also usually a terminal evaluation of a program that is already operational. Generally, there are two types of positivist designs; the true experimental design and the program group only designs. The second category is the interpretivism evaluation design, where evaluators respond to program participants and processes which are observed over time. As such, the interpretivism approach relies on the subjective association between the researcher and the subjects as well as the processes. The interpretevist approach is very much qualitative and process-oriented in nature. The third category is the mixed evaluation design, where the design draws upon elements from both the positivist and interpretevist paradigms. Therefore, it capitalises on and combines features from the summative and formative dimensions, as well as from the product and process dimensions. Finally, whichever design is selected, one needs to be wary of the design weakening factors.*

Keywords: evaluation, language program, design, positivist, interpretevist

1. INTRODUCTION

The future of language education is faced by a myriad of complex challenges posed by global and technological changes. In this regard, language programs are required to embrace contemporary dynamics in the field which cover a plethora of facets. These include developments in the field of English Language Teaching (ELT), the ever-changing global demand for language programs, the need to maintain quality amidst shrinking resources and to vie with more competitive providers. Therefore, the need to continuously improve existing language programs becomes an area that needs to be given serious attention. This is where the potential contribution of language program evaluation as a continuous improvement mechanism needs to be constantly explored and fully exploited as a means of ameliorating gaps in current language research and practice in ELT [1]. Much of the discussion here draws upon the works of [1, 2, 3,]. This paper looks at various evaluation designs available. It aims to shed some light on how potential language program evaluators can weigh in the various evaluation designs and apply what best suits them along the existing trends to maximize their undertakings. The rationale is that the conduct of language program evaluations can be fully exploited for the benefit of all stakeholders.

2. EVALUATION DESIGNS

[4] defined evaluation designs as "the conditions and procedures arranged by evaluators to collect data". Further to this, [3] explained the main keywords in the context of evaluation designs; "the treatment is the program, the experimental group is the program learners, and the control group is the group of learners to which the program learners are being compared". Among the factors to look into when considering the various evaluation design options, is the need to go back to the stakeholder requirements and purposes of evaluation [3]; what kind of information is required and why it is required. Only then an evaluator can effectively figure out "how" he or she is going to go about obtaining the information for future decisions. [3] mentioned and discussed the various evaluation designs as follows:

Positivist Evaluation Designs

Positivism is a paradigm that is generally quantitative in nature, where phenomena is measured and supported by objective evidence [5, 6]. It is at times also known as the scientific method. In evaluations, it usually involves the gathering of quantitative data or objective measurement. As explained by [2] and [7], the "quantitative dimension" it is usually summative in nature or product oriented. In this case, it is a terminal evaluation of a program that is already operational. Its purpose is to make judgments about a program's worth, its end result, or its effectiveness [7]. Being "scientific" in nature, it may involve the control or experimental group. [3] divided them into two broad categories, i.e. Comparison Group Designs and the Program Group Only Design.

The basic design to be preferred in this case is true experimental or quasi-experimental. In these situations, the program group learners receive a treatment (otherwise referred to as intervention, which is, in this context, the program instruction) while another group (the comparison group) receives nothing or it receives a different type of treatment. Both groups (if there are two groups) are measured using certain tests at different times throughout the program. The tests can be given before the start of the program (pre-test) and after the completion of the program (post-test). These can be either language proficiency tests whose content is related to the program curriculum in a general way, or they can be language achievement tests whose content is taken from the program curriculum [3]. He added that the designs for the experimental and quasi-experimental group are distinguished by three main factors:

a. Whether there is a control group or a comparison group.

b. How participants are assigned to a group (random or non-random).

c. The number of measurements taken (pre-tests, post-tests, time-series tests).

[3] described two types of designs for this category as follows:

a. Comparison Group Designs

[4] said that the use of control groups can eliminate most of the extraneous variables. The strongest of these designs is the true experimental design where learners are randomly assigned [4; 8]. However, according to [3] the opportunity for random assignment of participants is so rare that his literature concentrated mainly on quasi-experimental designs. Types of comparison group designs are as follows:

i. The True Experimental Design

This type of design involves a comparison group and random assignment of learners. Learners would also have to be randomly selected from the population of interest ([8]). It is the strongest design and the optimum way to ensure group equivalence ([4]). The main drawback is the fact that in numerous educational settings, randomisation is impossible [9; 8,3; 4]).

ii. The Classic Quasi-Experimental Design

This design involves a comparison between a program group and some other group of learners where the learners are not assigned randomly. Pre-existing differences between the groups can be adjusted using the "non-equivalent control group (NECG)". Measurements taken before and after the program for both groups can be statistically adjusted, as estimated by the pre-test. With this measure, the evaluation team can feel reasonably confident that the differences at the post-test stage are due to the program and not any other systematic differences between the groups [3].

iii. The Interrupted Time Series with Comparison Group

Here, periodic measurements are taken at several intervals before the introduction of the program, and after the completion of the program and the comparison group intervention. [3] said that the data gathered before the intervention (Time 1 through Time n) is used to predict what the data would look like if the same pattern continued after the intervention (Time n+1 through Time n+…). If the pattern of scores before and after the intervention showed differences, then the program has had an effect.

b. Program Group Only Designs

[3] said that the program group-only design is used in situations when it is difficult or impossible to find a suitable comparison group, or when only the program group is available for evaluation.

He also said that the program group-only design (also referred to by some as the pre-experimental or non-experimental designs) is weaker than the experimental or NECG designs. He described three types of program group-only design:

i. Program Group with Pre-test and Post-test

It is the same as the classic quasi-experimental design except that there is no comparison group. [3] said that although the conclusions that can be reached in this design are limited due to the lack of a comparison group, it still allows the evaluation team to have something to say about changes in learner learning and achievement throughout the program. He also said that the qualified conclusions from this design can be strengthened if the pre-test and post-test periods make use of multiple measures such as:

(1) Proficiency tests and achievement tests linked to the instructional goals of the program.

(2) Classroom observations.

(3) Questionnaires concerning perceptions of the program by teachers and learners.

ii. Longitudinal Designs

The longitudinal design, also sometimes referred to as "longitudinal study" ([10]) or "longitudinal survey" ([11]) is another design that can be useful when only the program group is available for measurement ([3]). Examples of this design are cohort studies or panel studies that study the sample from the program group over a period of time. In cohort studies different sample is used from the same population whereas in panel studies the same sample is used. In this type of design, learner achievements can be tracked from program records. The period of time of the study can be a semester, several semesters or several years [3]. Time available for evaluation is a disadvantage for this type of design as it can take many years to complete.

iii. Interrupted Time Series Design

It is the same as the quasi-experimental interrupted time series design except that it has no comparison group.

The main advantages of the quantitative evaluation design are the objectivity of the data (no bias) and the simplicity of analysis. The limitation is that it can only measure information that can be numerically expressed. It is unable to obtain in-depth information that involves subjective judgment or observation, such as the feelings and perceptions of program participants.

2. **Interpretevist Evaluation Designs**

The interpretevist paradigm is in contrast with the *a priori* nature of the positivist paradigm, which is *a posteriori* in nature. Here, the evaluators will respond to program participants and processes which are observed over time. The evaluators' interpretations and experiences are inherent throughout the process [3]. As such, in contrast to the numerical characteristics of the positivist approach, the interpretevist approach relies on the subjective association between the researcher and the subjects as well as the processes. These characteristics make the interpretevist approach very much qualitative in nature. We can now see that the interpretevist approach is very closely linked to the following evaluation dimensions; formative, qualitative and process [2, 7. 12] referred to qualitative measurement as "measurement of something that cannot be expressed numerically and that depends on subjective judgement or observation." One of the main reasons why this design has increasingly been incorporated into second language research is because of the difficulty in applying the controls necessary for experimental research in the classroom setting [13, 12] said that, generally, qualitative information is those obtained from classroom observations, interviews, journals, logs, and case studies. He also said that qualitative approaches are more holistic and naturalistic than quantitative approaches and seek to collect information in natural settings for language use and on authentic tasks rather than in test situations." The information is exploratory and large in amount gathered from a fairly small number of cases. The

information is usually open-ended and as a result, it is difficult to analyse. It must be coded or interpreted.

Others [3], said that this design "approaches a program as something to be observed and interacted with, rather than manipulated or measured." He also added that in this design, the evaluation team will "respond to program participants and program processes that they observe over the course of the evaluation, changing who they collect information from, when and how they collect it, as their understanding of the program develops." This, he said is in stark contrast with the fixed approach of the quantitative design. "This design involves, most importantly, an understanding of the program participant experience", said [3]. He added that there is also a degree of flexibility to this design in the sense that evaluators may use different data gathering techniques and instruments, or study different aspects of a program than originally planned as requirements and focus may change as events unfold during the evaluation. This technique also allows "progressive focusing' where evaluators can select particular aspects to look at in depth, and go back to the holistic view and so on ([3]). This can help them obtain a deeper understanding of the selected aspects and how they fit in the whole picture. The limitations to this design are subjectivity and the possibility of bias, time consumption, difficulty in analysis and interpretation of data and the high degree of engagement and presence required on the part of the evaluator.

Mixed Evaluation Designs

An evaluation can have a mixed-method (or "mixed strategies") where the design can draw upon elements from both the positivist and interpretevist paradigms. It can be a quantitative or qualitative one but uses data-gathering and analysis techniques from both [3, 5]. Therefore, it combines features from the summative and formative dimensions, as well as from the product and process dimensions. Leading experts such as [2, 3 ,12] argue for such evaluation designs that incorporate elements from these "opposing" dimensions for the following reasons:

a. Both types of methods serve different purposes and can complement each other [12].
b. The (multi-dimensional) interaction of audience, goals, context, and themes in an evaluation may suggest the need for such a design [2; 3].

[3] said that this design usually involves conducting an experimental-type inquiry and in-depth ethnographic-type inquiry at the same time. The qualitative element allows the evaluator to focus on certain aspects of interest of the program that could not be achieved by using the quantitative method alone.

The use of quantitative data can also be used to complement or validate the qualitative element. Therefore, this design allows the possibility of arriving at information that describes both measured effects of the objective view of the program as well as a multi-perspective view of the program ([3]). This design has numerous advantages [3], such as:

a. It can offer a richer set of information for decision making.
b. It allows the evaluators, even briefly, to "step outside" and view the program and its setting from a different perspective.
c. Even in the case of one of the designs holding sway, the resulting information can be more revealing than if only a single method is used.
d. Evidence from one method can help clarify findings from the other method.

However, [8, 3] also warned that this design may run certain risks, such as:

1. One of the designs may compromise the other.
2. It can result in contradictory findings – there is no guarantee that findings from both designs will triangulate around a single "truth" ([3]). He added that this can require reconciliation with other approaches to attain validity.

This method has relatively few disadvantages, which are; the length of time required for the evaluation, effort on the part of the evaluator to design instruments for both quantitative and qualitative designs, the effort and persistence to implement the studies and the relative difficulty in interpreting qualitative data and the need to collate both quantitative and qualitative data [3]. Despite this, the numerous advantages far outweigh the few disadvantages. To obtain more qualified conclusions for the purpose of making credible judgements, language program evaluations must incorporate multiple procedures for gathering information and multiple sources of information. This is because language program evaluations are complex and contain a diversity of features which require different types of information from different sources where a particular item of interest or phenomenon be explored in different ways and "triangulation' can be applied, where findings can be cross-checked across methods and sources for enhanced accuracy and validity [11; 14; 15; 3; 12; 16].

Others [15], said that the inclusive approach (mixed-method) to data gathering "provides different stakeholders with valid, credible and usable accounts". They even contrasted this method with other approaches and found other approaches as "narrower". [2] said that the use of both quantitative and qualitative data provides valuable information that should be used.

Recent developments strongly suggest the need for a mixed-method design for language program evaluations where both qualitative and quantitative methods can combine both summative/formative and product/process dimensions of evaluation [2; 11; 12; 15]. The gathering of as much information as possible from as many perspectives as reasonable will make an evaluation and the resulting decisions as accurate and as useful as humanly possible ([2]).

3. FACTORS THAT WEAKEN AN EVALUATION DESIGN

Having discussed the various evaluation designs, it must be borne in mind that any evaluation design runs the risk of being weakened due to several factors. Therefore, in selecting a particular evaluation design, an evaluator needs to be wary of these threats. [4] listed eight factors as follows that can weaken an evaluation design (these factors are also referred to as "extraneous factors"):

a. History – when instructional treatment extends over a considerable period of time, it is possible that other events may occur during that period that may have an additional effect on participants.

b. Maturation – natural growth (aging, maturation) of the participants during the program duration may contribute to the effects of the program treatment.
c. Testing – learners who are given the same test (as in the pre-test-post-test design) can become "test-wise" and perform differently on the post-test as a result of having taken the same test during the pre-test.
d. Instrumentation – if there is a change in the measuring instruments, then any changes in learner performance might be associated with the change of instruments other than the program treatment.
e. Instability – measures used in evaluation investigations can be not perfectly reliable and the resulting fluctuation of learner scores may erroneously suggest a treatment effect.
f. Selection – if learners from two or more comparison groups are selected differently for an evaluation study, the effects of the intervention can be confounded.
g. Mortality – if some learners from two or more groups in an evaluation study drop out differentially, the effects of the intervention can be confounded.
h. Statistical regression – if learners are selected for an evaluation study because of their extremely high or extremely low scores on a particular test, their performance on subsequent tests will tend to regress towards the mean of the distribution due to the statistical unreliability of the measuring device used.

4. CONCLUSION

From the discussion, it is evident that many leading experts view that the complexity and dynamic nature of language programs ideally call for mixed evaluation designs that can best address such concerns. However, having said that, each evaluation differs from the next one. Every evaluation has different goals, different audiences and different scopes ([17]). As such the value of a particular design must be carefully considered for it to arrive at the required outcome. The design must best suit the purpose of the evaluation, what kind of information is needed, whether there is a comparison group or a control group or not, as well as stakeholder requirements whilst being wary of the design weakening factors. These aspects will have to be weighed in accordingly to arrive at a design that best meets the evaluation needs with minimal compromise.

5. REFERENCES

[1] Norris, J.M. (2016). Language Program Evaluation. The Modern Language Journal, 100 (Supplement 2016), 168-189.
[2] Brown, J.D. 1989. Language program evaluation: a synthesis of existing possibilities. In Johnson, R.K. (ed.). The second language curriculum, pp. 222-241. Cambridge: Cambridge University Press.
[3] Lynch, B.K. 2003. Language assessment and program evaluation. Edinburgh: Edinburgh University Press.
[4] Popham, W.J. 1975. Educational evaluation. New Jersey: Prentice-Hall.
[5] Creswell, J.W. 2014. Research design. Qualitative, quantitative and mixed methods approach. Los Angeles: Sage.
[6] Fraenkel, J.R., Wallen, N.E. & Hyun H.H. 2012. How to Design and Evaluate Research in Education. New York: McGraw-Hill.
[7] Tufail, M.S., Husain, H.A., Salehan, D.A. & Azid, M.S. 2019. Devising an evaluation approach for a Malaysian armed forces english language program. Sci.Int.(Lahore),31(2), pp. 249-253.
[8] Lynch, B.K. 1996. Language program evaluation. Cambridge: Cambridge University Press.
[9] Beretta, A. 1986. Toward a methodology of ESL program evaluation. TESOL Quarterly. 20(1): 144-155.
[10] Gall, M.D., Gall, J.D. & Borg, W. B. 2003. Educational research. An introduction. Boston: Allyn and Bacon.
[11] Creswell, J.W. 2005. Educational research. New Jersey: Pearson.
[12] Richards, J.C. 2001. Curriculum development in language teaching. Cambridge: Cambridge University Press.
[13] Seliger, H.W. & Shohamy, E. 1989. Second language research methods. Oxford: Oxford University Press.
[14] Genesee, F. & Upshur, J.A. 1996. Classroom-based evaluation in second language education. Cambridge: Cambridge University Press.
[15] Kiely, R. & Rea-Dickins, P. 2005. Program evaluation in language education. Hampshire: Palgrave Macmillan.
[16] Weir, C. & Roberts, J. 1994. Evaluation in ELT. Oxford: Blackwell.
[17] Beretta, A. 1992. What can be learned from the Bangalore Evaluation. In Alderson, J.C. & Beretta Beretta, A. (eds.). Evaluating second language education, pp. 250-270. Cambridge: Cambridge University Press.