Low-Cost Pre-Evaluation of New Educational Programs

Bowen Hui, Bruce Hardy
Function Four Ltd.
Winnipeg, Canada
{bowen,bruce}@functionfour.ca

Yvonne Pratt
Communications
University of Calgary
Calgary, Canada
ypratt@ucalgary.ca

Rob Kershaw
Center for Digital Storytelling
Toronto, Canada
rob@storycenter.org

***Abstract*—Educational programs are developed to accommodate for new pedagogical findings and evolving curriculums. Methods to evaluate the effectiveness of these programs are typically labour intensive and time consuming — requiring the recruitment of program participants, the execution of the program, the collection of qualitative and quantitative data, and the analysis of results by expert researchers. To directly address the cost of such evaluations, we propose a pre-evaluation method that estimates the *expected value* of a new educational program before implementing it in practice. This approach allows educators, researchers, and stakeholders to obtain a preliminary assessment of new programs with minimal investment. To demonstrate our approach, we describe a case study that evaluates the impact of a digital storytelling workshop in a rural community.**

***Keywords*-program evaluation; decision theory; ICT skills**

I. Introduction

As new programs are introduced into the learning curriculum, systematic methods of evaluation are needed to assess the value of these programs with respect to their intended learning objectives. Educational programs are often evaluated through qualitative methods, such as case studies, content analysis, and grounded theory [1]. Quantitative methods adopted from behavioural psychology have also been applied to evaluate educational programs. One such approach is the use of pre-tests and post-tests; that is, through a pilot execution of the new educational program, students complete a skill test before and after the program so that changes in the test results (in comparison to the performance of a *control* group[1]) are credited toward the pilot program. In all of these cases, evaluation relies on data collected from executing the new program (in a pilot setting or in its full capacity).

Unfortunately, the execution of new educational programs comes with high costs. The evaluation study needs to be designed and conducted in a controlled and reliable way, so that the resulting data can be used to validate the effectiveness of the program. In particular, researchers and trained assistants are required to run the study, collect "clean" data, and analyze the results. Moreover, student participants are needed to undergo the new program in these evaluations. As a result, these methods are typically time consuming and labour intensive. In practice, evaluators are typically operating under limited resources — budget, turnaround time, and staffing. Thus, more cost-effective ways of program evaluations are needed to provide an early assessment of potential outcomes.

In this paper, we propose a simple *pre-evaluation* technique that directly addresses the cost of program evaluation. Our approach stems from decision theory in Economics and provides a normative evaluation of programs. We assume the availability of a standard assessment tool, such as a survey or an aptitude test, which we use as the benchmark to measure the current learning levels of the population of interest. Using this tool, we conceptually estimate the (hypothetic) change expected to be observed in the assessment of the population undergoing the newly proposed program. We use this information to obtain the *expected utility* of the program without actually executing it in practice. In this way, a program that has positive expected utility will likely yield an improvement in learning skills (in the overall population), according to the benchmark assessment tool. In contrast, a program that has an estimated non-positive expected utility is not worth further investment of time and effort. We elaborate on the details with examples below.

Our technique is particularly useful in decision scenarios where educators and funding agencies must choose a learning program for their schools or communities from multiple, available programs. By following the steps outlined in our approach, the stakeholders are able to assess the potential value of each program with respect to a pre-established benchmark. Thereafter, the program offering the highest expected utility according to these estimations would be chosen for further evaluation.

We emphasize that the purpose of our approach is to provide an early estimate of potential value in new educational programs before putting them in place. This work is not meant to replace existing evaluations; rather, it is designed to give an earlier, faster, and cheaper assessment. As such, this pre-evaluation technique compliments existing program evaluation methods. Moreover, it can be used in conjunction with any program evaluation methods.

[1]Students in the control condition take the same tests but do not participate in the pilot program.

The rest of this paper is organized as follows. Section II describes the pre-evaluation technique with illustrative examples. Our interest focuses on the community adoption of information and communication technology (ICT). As such, we describe a standard survey assessment for ICT called the *E-Index* in Section III. To demonstrate our method, we present a case study in Section IV, with emphasis on ICT skills. Lastly, we report the lessons learned in Section V.

II. Decision-Theoretic Program Pre-Evaluation

For comparison purposes, we assume that a typical program evaluation process consists of five steps as illustrated in Figure 1. In Step (1), the population of interest is identified and participants are selected (e.g., via a stratified sampling procedure [3]). In Step (2), participants take part in a pre-test that is deemed to be appropriate and sufficient for measuring the performance of the intended learning objectives of the pilot program. At this point, participants would be split into two groups: group A undergoes the pilot learning program (i.e., the *test condition*) and group B undergoes the regular learning program (i.e., the *control condition*). Generally speaking, the grouping of the participants should be done randomly, and in a way that allows the two groups to have (approximately) equal numbers. However, researchers may want to control for certain grouping variables, such as age, gender, and pre-test performance. In this case, we view groups A and B as having subgroups, and that the participants for each subgroup are selected randomly.

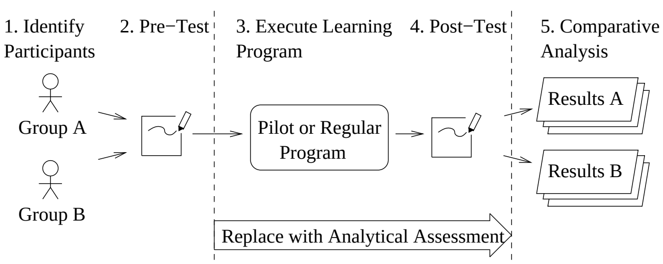

Figure 1. Process for program pre-evaluation.

Next, the learning programs takes place in Step (3), and the participants take a post-test in Step (4) upon completion. Variations of the pre-test may be used as a post-test, so long as the variables being measured by these tests are the same in both tests. Finally, in Step (5), the results are separated according to the original participant groupings and a comparative analysis (e.g., ANOVA [5]) is conducted to measure: (i) the performance difference between each group before and after the learning program, and (ii) the performance difference between the groups of the two programs.

In contrast, the pre-evaluation process that we propose replaces Steps (3) and (4) with an analytical calculation that estimates the projected post-test results. For example, if the assessment tool used is a survey, then an unbiased expert who understands the objectives of the pilot program needs to mine through each survey question, compare the pre-test scores obtained in the participant groups, and project a post-test score for each group. Moreover, if subgroups are used (e.g., group A male, group A female, group B male, group B female), then an estimated expected performance should be expressed for each subgroup.

We demonstrate the analytical assessment procedure with an example. Suppose a community's governing body is considering implementing a new technology training program that has demonstrated to be effective in other communities. How does one evaluate the effectiveness of such a program in this community without first executing it?

Table I
Example scores for two types of questions.

Q1	**Pre-Test Results**	**Post-Test Projections**
A (90/100)	30%	40%
B (75/100)	45%	50%
C (60/100)	10%	5%
D (55/100)	10%	5%
E (30/100)	5%	0%
Q2	**Pre-Test Results**	**Post-Test Projections**
Yes (1)	60%	75%
No (0)	40%	25%

To elaborate on the example, we suppose that a group of community members took a pre-test (e.g., survey) consisting of two types of questions with scores summarized in Table I. Suppose question $Q1$ is a type of question that assess performance and assigns a score between $0-100$ along with a summary letter grade. Through a hypothetical pre-test, we have 30% of the participants scoring a grade of A, 45% scoring B, and so on. On the other hand, if a question $Q2$ is of yes/no type (e.g., "do you know how to send emails?"), then we simply tally up the number of participants for each response in the pre-test. Obtaining pre-test results marks the end of Step (2) in the pre-evaluation process.

Next, we use the pre-test as a guide to estimate the maximal impact of a new educational program for this community. Going through each question and the participant scores one by one, a knowledgeable expert projects the expected change in the results if the training program were to be conducted followed by a post-test. For example, the pre-test evaluate one's telephone skills, computer skills, and Internet skills, while the new program promises to train participants on computer skills only. Thus, the skills that we would expect to see improvement in pertain only to computer knowledge. Example estimates for post-test scores are shown in the right-most column in Table I.

To calculate the expected change in a question, we calculate an average score for each scenario and subtract the difference. In particular, we use the median value for each score/response (e.g., 90 for A, 75 for B, etc.). An average score for the pre-test for $Q1$ is calculated by multiplying the percentage of participants who obtained each score and summing up all the possible cases as follows: $(30\% \times 90) + (45\% \times 75) + (10\% \times 60) + (10\% \times 55) + (5\% \times 30) = 70$. Following the same calculating using the projected post-test percentages, the average score is 79.25. Thus, the expected value of the learning program is $79.25 - 70 = 9.25$. One may further reference the grading scheme of the question to check whether this value improves the grade (say, from a B to an A). While such a number may seem abstract, ensuring that all the learning programs using the same benchmark assessment tool will enable a fair comparison in the same scale. An analogous calculation can be used for $Q2$, where a value of 1 is used for "yes" responses and a 0 otherwise.

III. E-Index: An ICT Assessment Tool

The current approach in community assessment is to conduct door-to-door surveys which are time consuming and labour intensive. Results in the quantitative components of these surveys are typically summarized as average participant responses. Here, we describe the community assessment survey called the E-Index that consists of questions about ICT adoption [4]. To date, the E-Index has been applied in 43 rural communities across Canada [2].

For a participating community, A housing list is used to establish the sampling frame. A random sample of households is drawn from this frame. As part of the rural community development initiative, local residents of the community are trained and certified as E-Index surveyors. These surveyors are responsible for conducting the survey with a member of their assigned households (as determined based on birth dates). Throughout the project, surveyors are supported by E-Index researchers remotely.

In total, there are four sections in the E-Index survey: demographics, ICT infrastructure, ICT skills, and ICT utilization. Example questions are: "Do you have access to $Tech$ at Loc", "Do you know how to use $Tech$", and "How often do you use $Tech$". where $Tech$ is replaced by radio, television, fixed phone, mobile phone, fax, computer, and Internet, and Loc is replaced by home, work, public areas.

Among other data collected in the E-Index, we focus on the quantitative results only. Responses for the questions in each section are averaged across all the respondents to obtain an infrastructure score, a skills score, and a utilization score for each of the seven technologies surveyed. These averages simply represent the percentage of respondents who indicated they have a technology at home, a skill for a technology, or a use for a technology. These percentages are then scaled to obtain a grade score using *goalposts* — a numeric score expressing the expected proportion of individuals who would indicate a positive response. In effect, if the goalpost is 100, the calculated percentage and the grade score are the same. However, if the goalpost is, say, 60, then only 60% of the population is expected to use the Internet. In this case, the grade score will be scaled to a number that is higher than the calculated (raw) percentage.

Although these scores are summary statistics, they may be used to help community leaders and policy makers understand their communitys adoption rates on various ICTs as a whole. For example, these scores may indicate that the infrastructure for Internet is very high but the actual skills to use it is very low. In this case, leaders may decide to invest in better training programs for Internet to increase knowledge and utilization of it. On the other hand, these scores may indicate that the community has very high skills in a technology but not enough infrastructure to support it. Such a result would suggest that leaders need to direct their investments to create broader access for that technology.

A. Assessment on Fishing Lake, Alberta

In 2008, Fishing Lake Métis Settlements participated in the E-Index project (version 2.4). A total of 148 households were used to establish the sampling frame and 84 households were sampled. With a response rate of 89.3%, a total of 75 surveys were completed successfully.

B. Project Findings

The overall E-Index score for Fishing Lake is 64.5% (a letter grade of B). Table II shows a further breakdown of the category and technology grade scores. In comparison to 17 other communities who participated in this version of the E-Index, Fishing Lake has similar scores in about half of the cases, with Infrastructure, Utilization, Fixed phone, Mobile phone, and Computer scoring slightly below the average.

Table II
E-Index scores for Fishing Lake in 2008.

Category/Technology	Grade Score	e2.4 Average
Infrastructure	79.7 (B)	88.7 (A)
Skills	91.6 (A)	90.1 (A)
Utilization	27.6 (D)	52.4 (C)
Radio	83.5 (A)	84.8 (A)
Fixed Phone	87.8 (A)	92.8 (A+)
Fax	44.4 (C)	52.3 (C)
Mobile Phone	68.8 (B)	79.6 (A)
Television	74.0 (B)	80.1 (A)
Computer	50.8 (C)	59.9 (B)
Internet	42.1 (C)	55.8 (C)

IV. Case Study: Digital Storytelling

We report our experience with a pilot study in Fishing Lake that was conducted in the Fall of 2010. This study is centered around a digital storytelling (DST) workshop, where DST experts are invited into the community to train youths on various technologies and teach them about digital story making. After a period of hands-on training, the participants are required to create a digital story in teams. In this study, seven types of software and hardware were used in the digital storytelling workshop. These technologies are:

- Word processing software (WP): MS Word, Notepad
- Movie editing software (ME): Final Cut Studio, Windows Movie Maker

- Storage device (SD): USB key, CD, DVD
- Digital camera (DC)
- Camcorder (CC)
- Scanner (SC)
- Audio recorder (AR)

To assess the specific technology skills, we extended the E-Index questionnaire with detailed questions pertaining to the knowledge (with yes/no questions) and utilization of these technologies. The objectives of the workshop is to expose new technology to participants, equip them with the necessary skills to use the technology, and inspire them to explore technology use in their daily lives.

A. Calculating the Expected Impact

A total of 5 youths participated in this pilot project. Participants completed the extended E-Index questionnaire as the pre-test. Results are shown in Figure 2. With emphasis on the technology skills questions only, Table III shows the corresponding distributions and average technology scores.

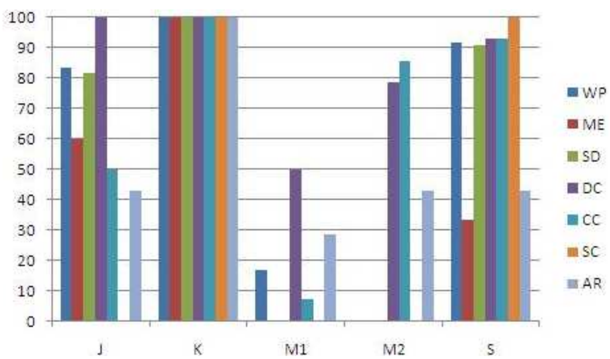

Figure 2. Pre-test results.

Based on the participants' engagement level, the complexity of the technology, and their perceived enjoyment level (as reported by the participants), we augmented the distributions and calculated the projected averages shown in the last row of Table III. Averaging across the seven technologies, we expect an improvement of 14.3 points from the workshop.

Table III
EXAMPLE SCORES FOR TWO TYPES OF QUESTIONS.

Grade	WP	ME	SD	DC	CC	SC	AR
A	0.6	0.2	0.6	0.6	0.6	0.4	0.2
B	0	0	0	0.2	0	0	0
C	0	0.2	0	0	0	0	0
D	0	0	0	0.2	0.2	0	0
E	0.4	0.6	0.4	0	0.2	0.6	0.8
Pre-Test Averages	66	48	66	80	71	54	42
Projected	81	74	76	87	84	63	62

B. Preliminary Results

About two months later after the workshop was over, participants were asked to fill out the E-Index again as a post-test. To minimize the effort in answering the same questions twice, we created an online survey tool that automatically populated the post-test responses using the responses provided from the pre-test. Post-test results (actual scores) are shown in Figure 3.

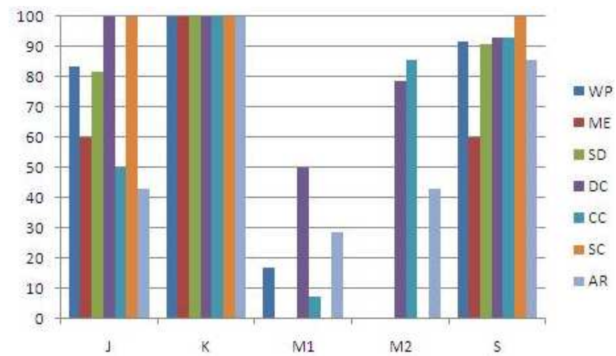

Figure 3. Post-test results.

Due to the small number of participants and a lack of a control group, we cannot conduct rigorous analyzes on the quantitative results to measure the impact of the workshop nor to assess the value of our pre-evaluation process. Through casual observations, community members noticed that after the workshop, the student participants began to spend more time in the technology laboratory at the community centre.

V. Conclusion

We presented a simple pre-evaluation process to address the cost of program evaluations. This technique is designed to compliment existing program evaluation approaches with an earlier, faster, and cheaper estimate assessment of new programs. Through analytical examples and preliminary data, we demonstrated the details of the technique. Further empirical evidence is needed to assess the value of this pre-evaluation process.

References

[1] D. Ary, L.C. Jacobs, C. Sorensen, and A. Razovieh. (2010) *Introduction to Research in Education.* Wadsworth Publishing.

[2] B. Hui, S. Irwin, and P. Robins. (2010) *Cultural Community Comparison: First Nations, Métis Settlements, and Non-Aboriginal Communities.* DVA reports, Function Four Ltd.

[3] L. Kish. (1995) *Survey Sampling.* Wiley-Interscience.

[4] W. Kelly. (2008) *Knowledge Planning: Community Development in the Knowledge-Based Economy.* Department of Rural Development, Brandon University, Brandon, Canada.

[5] P.S. Maxim. (1999) *Quantitative Research Methods in the Social Sciences.* Oxford University Press.