*Qualitative and Quantitative Methods in Libraries (QQML) 3: 293 – 303, 2013*

# Does Evaluation Improve Performance? – A Case Study of Japanese Public Libraries –

**Kanako Shimoyama[1], Minako Nishiura[2], Hajime Naka[3], Keita Tsuji[4], and Hiroshi Itsumura[5]**

[1, 2, 3] Graduate School of Library, Information and Media Studies, University of Tsukuba
[4,5] Faculty of Library, Information and Media Science, University of Tsukuba

**Abstract: Purpose –** This study explores the effectiveness evaluation program in public libraries by conducting empirical analysis of selected Japanese public libraries.
**Design/Method/Approach –** We used three methods: 1) the analysis of libraries' performance data; 2) meta-evaluation; and 3) case study.
**Findings –** The evaluation program in the public libraries does not necessarily contribute to the improvement of the service performance. Additionally, the type of the evaluation (performance measurement or program evaluation) had certain effects on the performance improvement.
**Originality/value –** This study provides new insight into on-going researches on evaluation program in libraries.

**Keywords:** Library management, library administration, performance measurement, public libraries, self-evaluation, program evaluation, meta-evaluation, evaluation studies

## 1. Introduction

### 1.1 Background

As of 2011, there were 3,196 public libraries established and managed by local governments across the country in Japan[1]. Each public library is required by the Library Law to make efforts to 1) conduct evaluation programs, and 2) implement improvement measures.

### 1.2 Literature Review

[1] Of those, 61 were operated by prefectural government; 223 were operated by the special wards of Tokyo; 2,321 were operated by other cities (of those, 276 were operated by government ordinance-designated cities); 590 were operated by towns and villages; and 1 was operated by a large municipal area.

Received: 16.5.2013 / Accepted: 20.8.2013 ISSN 2241-1925

According to the 2008 report from Mizuho Information and Research Institute, Inc. (MIRI, 2009) of a library survey commissioned by the Ministry of Education, Culture, Sports, Science & Technology in Japan: 1) of the total valid responses (1,772 libraries), 21.0% (373 libraries) were conducting self-evaluations[2]; 2) of those, 366 libraries answered the question about problems with the evaluation program, and of those, 57.1% (209 libraries) believed that the problem involved practically implementing the evaluation results; 3) also among of the total valid responses, 371 libraries answered the question about how they use the evaluation results, and of those, 79.5% (295 libraries) responded that they use the results as basic material for management.

Less attention has been paid to the confirmation and verification of the actual effectiveness of evaluation programs in Japanese public libraries. For instance, in 2008 the National Council of the Public Library in Japan (NCPL) conducted a questionnaire survey to ask public libraries conducting self-evaluation programs whether they use the results from the program. Although NCPL's study reveals to what extent respondents use the results for performance improvement, it gives us no evidence of the effectiveness of the evaluations because it failed to ask the respondents whether or not their evaluation programs succeeded. Outside of Japan, Poister (2010) points out that there is no comprehensive research on evaluation program in terms of particular type of evaluation. Thus, there is not much evidence that evaluations actually have positive effects on performance improvement. This study therefore aims to examine to what extent evaluation programs in public libraries contribute to and positively affect performance improvement, and analyze those findings.

### 1.3 Definition

#### Evaluation

Although the term "evaluation" has been defined in many ways, we define it as "evidence of the effectiveness and accomplishment of a particular policy, management, or operation," which comes from Furukawa (2001).

#### Program and Performance

This study adopts the definition given by the U.S. Government Accountability Office, as well as its former organization, the U.S. Government Accounting Office (GAO). According to GAO (2011), a program was defined as "any activity, project, function, or policy that has an identifiable purpose or set of objectives." GAO (1992) also defines performance as certain collective data including the program's: 1) input, such as dollars, staff, and materials; 2) workload or activity levels; 3) output or final products; 4) outcome of products or services; and 5) efficiency... In order to examine how evaluations in public libraries positively affect performance improvement, this study used three measurements: the number of reference service use, the number of library

[2] Of the total valid responses (1,772 libraries), 48.0% (856 libraries) conducted evaluation programs managed by the local government, 30.6% (543 libraries) did not conduct any evaluation programs, including self-evaluation.

visitors, and the number of circulation of materials, which corresponds to the fourth part of the definition by GAO above.

**Performance Measurement and Program Evaluation**

Furukawa (2001) and GAO (2011) state that the evaluation methodology can be categorized into two types: performance measurement and program evaluation. This study defines performance measurement as the process of comprehensive analysis of the whole workflow of the organization based on a selected measurement (or data, or variable). The number of circulation of materials can be counted as this type of measurement. This study also defines program evaluation as the process of in-depth analysis of an instance of policy and operations using methodologies employed by the social sciences such as interview.

### 1.4 Hypothesis

Our hypothesis is that library evaluations with performance measurement alone tend to fail to improve performance.

## 2. Methods and Materials

In our research, we used three methods: analysis of performance data, meta-evaluation and case study. This chapter describes the outline, sampling, and the details of each of the three methods.

### 2.1 Outline

First, we analyzed the performance data to examine whether library evaluations improve performance based on three performance indicators: the number of reference service use, the number of visitors, and the number of circulation of materials. For this analysis, we used the data[3] in *Statistics on Libraries in Japan (SLJ)*, the annual monograph published by the Japan Library Association.

Secondly, we tested our hypothesis, using both quantitative and qualitative data. We collected the former data through meta-evaluation, grading the evaluation reports according to how clearly needs or improvements are described. We collected the latter data thorough case studies comprising interviews and literature research.

### 2.2 Sampling

We focused on libraries actively conducting evaluations. As MIRI (2009) reported that 11 public libraries are actively conducting evaluations. Additionally, NCPL (2010) reported 16 cases. Although NCPL did not explain what those cases meant, four cases overlapped with the data from MIRI (2009), which made it possible to conclude that the meaning in both cases was the same. The research reports comprise 23 non-overlapping cases. We obtained 16 libraries using our criteria: the library evaluated itself of its own initiative; and the report was made public with information pertinent to our research theme. Selected sixteen libraries were alphabetized from A to P.

---

[3] The reference service use and circulation material data were from 1995 to 2011, and the visitor number data were from 2003 to 2011, because the Japan Library Association only collected the latter from 2003.

**Type of Evaluation**

We investigated the type of evaluation adopted by selected 16 libraries using their evaluation reports (Table 1). The result shows that 14 libraries used only performance measurement: comprehensive analysis of the whole workflow of the organization based on a selected performance indicator and reporting that performance. Besides performance measurement, the remaining two libraries conducted their evaluations using several methods: interviews, analysis records of reference service use, user surveys, text mining and analysis the health and medical information service and web-based library projects. Thus we see those two libraries conducted program evaluation.

Table 1. Selected libraries (n=16)

| Library | Municipality | Type of Evaluation |
|---|---|---|
| A | Prefecture | PM |
| B | | PM |
| C | | PM |
| D | | PM |
| E | | PM, PE |
| F | | PM |
| G | | PM |
| H | | PM |
| I | | PM |
| J | Capital Ward | PM, PE |
| K | | PM |
| L | Ordinance-designated City | PM |
| M | | PM |
| N | City | PM |
| O | | PM |
| P | Town/Village | PM |

PM: Performance Measurement, PE: Program Evaluation

**2.3 Analysis of Performance Data**

We analyzed the performance data to examine whether library evaluations improve performance. We learned when selected libraries chose certain indicators, and measured performance through three sources: MIRI (2009), NCPL (2010), and the websites of the selected libraries. If we could not obtain the necessary information, we inquired with the library directly.

**Performance Indicators**

To identify common trends across the libraries in performance level increase/decrease per performance indicator associated with evaluation, we chose a much larger sample size as soon as possible. We first examined the list

of indicators based on the history of evaluation reports published by the libraries on their websites. Some libraries listed more than 60 items as performance indicators. Of many applicable performance indicators in *SLJ*, three were found to be the top three most used indicators among our samples: reference service use (11 libraries); visitors (7 libraries); and circulation of materials (6 libraries). We then examined performance level per indicator of the libraries that listed one, two or all three of these indicators. For instance, if the library listed reference service use as a performance indicator, but did not list the other two, we examined reference service use only.

Although the range of the period of data collection in *SLJ* was from 1995 to 2011, we limited our scope to 1995 to 2010, and omitted cases where the span of operations for each indicator was too short to enable a focus on long-term performance trend. (Some libraries began to use certain indicators after 2010; some stopped using certain indicators halfway through period covered by *SLJ*. Such cases were omitted). After modifying the period of data collection, we finalized our selection of three performance indicators: reference service use (10 libraries); visitors (6 libraries); and circulation of materials (5 libraries).

**Methods for Analyzing the Performance Data**

Data were derived from *SLJ* from libraries using one, two, or all three of our indicators: reference service use; visitors; and circulation of materials. In terms of the average performance level score before, during and after evaluation was conducted, we conducted an *F*-test first to check for homoscedasticity, and then conducted a *t*-test based on the results of the *F*-test. We set the significance level to $p < 0.05$, and analyzed trends in performance level increase/decrease. For instance, the average score of the number of reference service use at Library A significantly increased after beginning to use this indicator as an evaluation of performance level compared to the average before 2008 ($p = 0.001$).

To verify whether our selected libraries' performance level trends also apply to other libraries, we conducted a *t*-test for other libraries using the same procedure. To ensure fair judgment, we first selected libraries managed by the same local governmental bodies as the libraries we selected from the data in *SLJ*. We then calculated mean population per fiscal year for each performance indicator: reference service use; visitors; and circulation of materials. For the purpose of fair estimation of to what extent and from what point the other libraries' performance level increased or decreased compared to that of our selected libraries, the average of the fiscal year of each performance indicator from our selected libraries was used as a borderline. Table 2 shows the fiscal year of the first evaluation and its average for each of our selected libraries per indicator.

**2.4 Meta-evaluation**

For the purposes of this study, meta-evaluation was taken to mean an evaluation of how clearly improvements or needs are described by use of a checklist. First, we applied the "Recommendations and Explanations" section

Table 2. Fiscal year of the first evaluation of each library and indicator

| Municipality | Library | Indicator | | |
|---|---|---|---|---|
| | | Reference Service Use | Visitors | Circulation |
| Prefecture | A | 2008 | | |
| | B | 2003 | | |
| | C | 2006 | | |
| | D | 2005 | 2008 | |
| | E | 2008 | 2008 | |
| | F | 2001 | 2004 | 2001 |
| | H | 2006 | 2006 | 2006 |
| | I | 2008 | 2008 | 2008 |
| | Avg. | 2005 | 2006 | 2005 |
| Ordinance-designated City | L | | 2006 | |
| | Avg. | | 2006 | |
| City | M | 2008 | | 2008 |
| | N | 2006 | | 2006 |
| | Avg. | 2007 | | 2007 |

(the focus of our research) of the "Key Evaluation Checklist" made by Sasaki (2008) for aid evaluation work. Table 3 shows the scores and meanings. The lowest mark was zero, and the highest was four. Although Sasaki (2008) used evaluation reports generated in a single fiscal year, we could not choose the year, because the selected libraries in our research did not publish reports every year. Therefore, evaluation reports were chosen from the latest evaluation report that had been published by each library as of June 2011.

Secondly, we divided sixteen libraries into Group I and Group II based on the type of evaluation they conducted. Group I conducted performance measurement only, and Group II conducted both performance measurement and program evaluation. Thirdly, after scoring each library's evaluation report, we calculated the average score for Groups I and II. Finally, we made a comparison of those two average scores.

Table 3. Scores and Meaning

| Score | Meaning |
|---|---|
| Excellent (4) | All (or almost all) listed items are covered and appropriately examined |
| Good (3) | Most listed items are covered and appropriately examined |
| Satisfactory (2) | Most listed items are covered and briefly examined |
| Weak (1) | Some items are covered and briefly examined |
| Unsatisfactory(0) | Almost (or completely) no listed items are covered |

Source: Sasaki (2008)

**2.5 Case Study**

To conduct a comparison study on the status of management and service operations between libraries that used different evaluation results to improve, we selected the following two libraries: Library F, which conducted performance measurement only, and Library J, which conducted both performance measurement and program evaluation. We chose Library F for several reasons. First, they started to work on self-assessment at a relatively early stage and still have a positive attitude toward it. Second, they have published quite a few reports on evaluation activities in their own bulletin and some other journals. As for Library J, we selected it because of their adoption of social science methods that met the definition of program evaluation as shown in the preceding chapter.

We interviewed two people with a good knowledge of each library's evaluation. The first author of this paper conducted two semi-structured interviews. One was a one-on-one in-person in July 2011 with a librarian who worked for Library F, 75 minutes in duration. The other was an interview via series of e-mails (sent between December 2011 to January 2012) with the chairman (at the time) of the evaluation sub-committee of Library J council.

## 3. Results

**3.1 Analysis of Performance Data**

Tables 4-6 present performance trends for three indicators around the time when the libraries started to refer to their performance in their evaluation reports. In Tables 4-6, performance increase or decrease is represented by a plus sign (+) or a minus sign (-), and (ns) is standing for not significant.

Table 4. Trends in Reference Service Use

| Municipality | Library | Trend | *p* |
|---|---|---|---|
| Prefecture | A | + | <.001 |
| | B | + | <.001 |
| | C | − | .001 |
| | D | − | .03 |
| | E | − | .02 |
| | F | − | .04 |
| | H | + | .03 |
| | I | ns | .62 |
| | Parameter | ns | .39 |
| City | M | − | .004 |
| | O | + | .009 |
| | Parameter | ns | .87 |

Table 5. Trends in Visitors

| Municipality | Library | Trend | *p* |
|---|---|---|---|
| Prefecture | D | ns | .38 |
| | E | ns | .93 |
| | F | − | .02 |
| | H | ns | .07 |
| | I | ns | .15 |
| | Parameter | + | .03 |
| Ordinance-designated City | L | − | .004 |
| | Parameter | ns | .18 |

Table 6. Trends in Circulation

| Municipality | Library | Trend | *p* |
|---|---|---|---|
| Prefecture | F | + | <.001 |
| | H | ns | 2.40 |
| | I | ns | .05 |
| | Parameter | + | <.001 |
| City | N | ns | .51 |
| | O | ns | .30 |
| | Parameter | ns | .14 |

**3.2 Meta-evaluation**

Table7 shows of meta-evaluation result. The average score was rounded to one decimal place. The average score of Group I was 2.1, and Group II was 4.0; those results showed a significant difference ($p < 0.001$). These results indicate that the former group failed to clarify the points of improvement they should have made in their reports.

Table 7. Meta-evaluation Scores

| Group | I | | | | | | | | | | | | | | II | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Library | B | C | D | F | G | H | I | K | L | M | N | O | P | Avg. | E | J | Avg. |
| Score | 3 | 2 | 0 | 3 | 0 | 2 | 3 | 4 | 0 | 4 | 3 | 1 | 2 | 2.1 | 4 | 4 | 4.0 |

*Note*: As Library A stated they would evaluate their performance in 2011, they had not yet implemented it when they published their report. Hence, we did not give it a score, and excluded it from the denominater.

### 3.3 Case Study

We describe the findings from our interviews and document investigation. First, we show the major characteristics of Library J: adopting a designated administrator system. Library F did not have such a system. Secondly, Library F, which employed the performance measurement method alone, has not been able to address and improve the issues of the organization as a whole. Furthermore, their satisfaction with simply identifying the problems prevented them from thinking about how to improve the situation. Thirdly, on the other hand, Library J, which used both performance measurement and program evaluation methods, has improved the operational framework of the entire organization. Additionally, they have developed a mechanism for reassessing unsolved problems years after year until an improvement measure is found.

## 4. Discussion

As mentioned in the introduction, the aim of this study is to reveal and clarify the effectiveness of library evaluation programs. Our findings from the analysis of performance data indicate that while performance levels in some libraries have significantly increased after beginning to conduct evaluations, they have significantly decreases in some libraries, which leads us to the conclusion that evaluations do not necessarily contribute to performance improvement. Unfortunately, our findings do not explain the association between the practice of evaluation and the range of increase/decrease in performance level. Further analysis is needed to clarify the contradictory factors that cause variations in performance level after evaluation.

On the other hand, the results from our meta-evaluation checklist analysis, calculated in order to learn to what extent evaluation reports cover the items listed for modification of service performance, show that libraries adopting only performance measurement score lower than libraries adopting both performance measurement and program evaluation. We consider this result to be significantly useful for supporting our hypothesis, because the result implies that the higher the library's score, the more they can identify what needs to be improved, and heed the importance of performance improvement.

Case study reveals that while libraries adopting only performance measurement are more likely to limit their improvement efforts, libraries adopting both performance measurement and program evaluation are more likely to successfully solve their problems. We conclude that a two-pronged methodological approach facilitates the performance improvement.

We found that the meta-evaluation and case study findings support our hypothesis that choice of evaluation methodology can contribute to and positively affect performance improvement.

**Limitation in our Methodology**

This study has limitations. One limitation is the relatively small sample size, especially for case study. Thus, the findings may not apply to other cases or other samples. For instance, this study considered the difference in choice of evaluation methodology as a factor in performance level, but did not consider other differences such as the presence or absence of participation in designated administrator systems (as in the case of Library J and Library F), which might also be factors that positively affect improvement efforts. The other limitation is incomplete evaluation report data. Because it is voluntary, some libraries choose not to disclose certain management situation and evaluation criteria information. In those cases, there is no information available for such libraries, but it does not mean that such libraries do not conduct evaluations.

## 5. Conclusions

In conclusion, we found from our examination of performance data that evaluations do not necessarily contribute to performance improvement. We also found from meta-evaluation and case study analysis that library evaluations adopting only the performance measurement method tend to fail to improve the library service performance. Further study with a much larger sample size will be needed in order to generalize the effectiveness of using both the performance measurement and program evaluation methods, applicable to many Japanese public libraries.

**Acknowledgements**

This research owes much to the thoughtful comments of Professor Hideki Minai. The authors thank Professor Mitsuhiro Oda (Aoyama Gakuin University) and the librarian at Library F who were kind enough to share their insights with us through interviews. This work was supported by Research Center for Knowledge Communities University of Tsukuba.

**References**

Furukawa, Shun'ichi., (2001). *Kokyo Bumon Hyoka no Riron to Jissai*, Nihon Kajo Shuppan, Tokyo. (text in Japanese).

Japan Library Association., (1995-2011). *Statistics on Libraries in Japan*, Japan Library Association, Tokyo. (text in Japanese).

Mizuho Information & Research Institute, Inc., (2009). *Tosyokan no Jiko Hyoka Gaibu Hyoka oyobi Un'ei no Jokyo ni Kansuru Joho Teikyo no Jittai Chosa Hokokusyo*: *A Study Commissioned by The Ministry of Education, Culture, Sports, Science and Technology*, Mizuho Information & Research Institute, Inc., Tokyo. (text in Japanese).

National Council of Public Libraries., (2009). *Koritsu Tosyokan ni okeru Hyoka ni Kansuru Jittai Chosa Hokokusyo*, National Council of Public Libraries, Tokyo. (text in Japanese).

National Council of Public Libraries., (2010). *Koritsu Toshokan ni okeru Hyoka ni Kansuru Hokokusyo*, National Council of Public Libraries, Tokyo. (text in Japanese).

Poister, Theodore H., (2010). *Performance Measurement.* Wholey, Joseph S., et al, (ed.) *Handbook of Practical Program Evaluation.* 3rd ed, 100 – 124.

Sasaki, Ryoh., (2008). *Metaevaluation by Formal Evaluation Theory of Aid Evaluation Work*, PhD thesis. Western Michigan University.

The U.S. Government Accountinig Office (GAO)., (1992). *Program Performance Measures: Federal Agency Collection and Use of Performance Data.* Report GAO/GGD-92-65. Washington DC.

The U.S. Government Accountability Office (GAO)., (2011). *Performance Measurement and Evaluation: Definitions and Relationships.* Glossary GAO-11-646SP. Washington DC.

.