

Understanding the Methodological Foundations of  
Public Library National Rating Systems

by Ray Lyons, MLIS, MPA  
Independent Consultant  
Cleveland, Ohio, USA

Paper Presented at the IFLA Statistics and Evaluation Section  
Satellite Post-Conference Meeting of the  
74th IFLA General Conference and Council

Concordia University  
Montreal, Québec, Canada

August 18-19, 2008

## Standardized Statistics in the History of Statistical Ideas

The global library statistics model developed by the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute of Statistics, International Organization for Standardization (ISO), and International Federation of Library Associations and Institutions (IFLA) promises to provide a wealth of information about libraries and their contributions to their nations and societies. As statistical historian Desrosières reminds us, ambitious 21<sup>st</sup> century projects such as this one are built upon statistical traditions that have developed over several centuries.<sup>1</sup> Arguably, the rationale for global statistics is based directly upon ideas from 17<sup>th</sup> and 18<sup>th</sup> century practices in German *descriptive statistics* and English *political arithmetic*.

Desrosières notes two milestones in the progression of statistical thinking that are pertinent to the collection and use of library statistics, and to making comparisons with these data. These milestones are the *creation of equivalences* and *encoding*. Creation of equivalences is the establishment of standard classifications to describe phenomena relevant to a nation or society—persons, groups, events, objects, industries, institutions, jurisdictions, and so on. The classification scheme emphasizes similarities between the phenomena and ignores their numerous differences. Encoding is the specification and use of definitions to assign the various phenomena to the classifications.

Library statistics and performance indicators are obvious derivative applications of these historical ideas. These foundational concepts are sources of both the strengths and weaknesses of standardized statistical data collection. Using traditional library statistics in the aggregate, as public library national ratings do, exacerbates the weaknesses inherent to these statistical collections. To a lesser extent, these weaknesses also affect more routine statistical comparisons of individual libraries. This paper will explore how the creation of equivalences and other characteristics of library statistical data reduce the accuracy and reliability of comparisons of library performance.

### National Library Ratings Systems

Among approaches to comparing library operational statistics, public library rating systems are distinguished by their use of composite statistical scores. The rating systems presume that a summary depiction of library performance is desirable and can be accomplished by combining individual performance indicators into singular scores.

In the USA, national public library statistics are collected via the Public Library Statistics Cooperative (PLSC). Initially called the Federal-State Cooperative System, PLSC was formed in 1980 as a collaborative effort by the U.S. Department of Education, U.S. National Commission on Library and Information Science, American Library Association (ALA), state library organizations, and others. The cooperative began publishing national statistical data on an annual basis in 1991.

In 1999, American library consultant Thomas Hennen used PLSC data to create proprietary ratings of public libraries known as the Hennen American Public Library Ratings (HAPLR).<sup>2</sup> Hennen has produced these ratings annually and ALA published them each year until 2007. In its first two years HAPLR inspired considerable controversy among American libraries. Critics noted the small number of performance indicators used and questioned the theoretical basis for the calculation formulae. Others (including highly-rated libraries) supported HAPLR as a sound and reasonable evaluation approach. In the ensuing years the debate subsided and the annual ratings have been published annually—unabated and unchanged.

While serving as a graduate intern at the U.S. National Commission of Library and Information Science, I conducted an in-depth study of the HAPLR methodology. In the study I noted that the ratings have been faulted for lacking a clear explanation of what they were intended to measure.<sup>3</sup> I also suggested that HAPLR's intricate calculations made it impossible for libraries to determine the exact criteria by which they were being rated. Further, the rating system failed to include a clear and consistent account of the limitations of the methodology. Nor had the HAPLR system included appropriate guidance in interpreting ratings in a manner consistent with the methodology and data used.

With my colleague Keith Curry Lance (former director of the Library Research Service in Colorado) I have recently participated in the design of a new American public library national rating system called the *LJ* Index.<sup>4</sup> Instituted by the *Library Journal*, the first edition of these ratings will be issued later this year. Central to *LJ* Index is a program of education aimed to assure that libraries understand the ratings and their methodological foundations. We do this to impress upon the library community that, regardless of their designs, library ratings are necessarily simplistic and unsophisticated assessments of library performance.

### Performance Measurement Ideology

Library ratings systems are based on traditional library statistical indicators that have been promoted as useful for assessing public library performance and effectiveness.<sup>5</sup> The data are also thought to reflect quality and value.<sup>6</sup> Yet, there has been growing dissatisfaction with library *enabling* (input) and *use* (output) statistics. For instance, Hernon and Altman concluded that these statistics are inconsequential because they:

...do not measure the library's performance in terms of elements important to customers. They do not really describe performance or indicate whether service quality is good, indifferent, or bad.<sup>7</sup>

Indeed, limits to the meaning and substance of traditional enabling and use measures have led to the pursuit of more convincing measures of library outcomes, impacts, and value.

Nevertheless, the main rationale for gathering information of either type—enabling and use measures, or outcome and impact measures—comes from the tenets of *performance management*.<sup>8</sup> This management approach adopts a rational view of organizations and advances the collection of data that will be “actionable.” Theoretically, operational statistics provide valuable feedback that will contribute directly to improved decision-making, which, in turn, will improve organizational performance. The approach, also known as *results-oriented management*, is central to literature on library assessment as well as to performance measurement literature in public administration, program evaluation, business excellence, and quality management.<sup>9</sup>

In the USA in the mid-1980’s, results-oriented management was central to the program of standardized library statistics promoted by Public Library Association. This program argued that use (output) measures “reflect results or outcomes, the effectiveness and the extensiveness of the services delivered by the library.”<sup>10</sup> Well-managed libraries were expected to track the magnitudes of use statistics and to consider these data as legitimate *results* that library managers and stakeholders would use to monitor the progress and effectiveness of libraries. In practice if not in theory, the idea that library use statistics are synonymous with results is the predominant view among American public libraries today.

Incidentally, it is important to note that, despite claims made by its promoters, the effectiveness of results-oriented management has not been demonstrated. Cullen has questioned whether performance measurement actually leads to performance improvement.<sup>11</sup> Radin notes that the approach makes unrealistic and inappropriate assumptions about how organizations function.<sup>12</sup> And both Spitzer and Grizzle recount how “dysfunctional” measures commonly found in public and private sector organizations produce a variety of negative and unintended consequences.<sup>13</sup>

### Using Comparative Library Statistics

A practice central to results-oriented management is benchmarking, the use of comparative data from similar organizations to assess the performance of one’s own organization. This is a fundamental tool in both quality management and performance scorecard approaches. This tool has also been enthusiastically advanced to local governments in the form of *comparative performance measurement*.<sup>14</sup>

The primary reason for acquiring comparative data is the lack of objective criteria by which local governments, libraries, and other organizations can evaluate their own performance data. Comparative data ostensibly help to “place local performance in context and, where major performance gaps are detected, may suggest the need for additional analysis.”<sup>15</sup> This approach has been promoted as essential to library management and assessment.<sup>16</sup>

Even so, library comparisons are neither straightforward nor necessarily conclusive. Poll and te Boekhorst are careful to provide numerous caveats about making comparisons

using statistical indicators.<sup>17</sup> They make repeated admonitions about interpreting comparative findings cautiously by looking for alternative explanations for measurement variances. Other proponents of comparative performance assessment acknowledge the fact that these statistics should be viewed with a certain amount of skepticism. For example, Ammons notes that localities having high performance statistics may still be neglecting particular constituent populations, and that local statistics are self-reported, unaudited, and susceptible to errors.<sup>18</sup>

A more troubling problem with these comparisons is the imperfect methods for selecting peer organizations. Morely, Bryant, and Hatry conceded that “no two . . . jurisdictions or organizations are completely comparable. Each has unique characteristics. As a result, it is impossible to find organizations that are exactly comparable.”<sup>19</sup>

In other words, we currently have no tools for accurately measuring organizational comparability. For this reason, benchmark comparisons, at their best, will be gross estimates. At their worst, they will be inaccurate and misleading. Presently, the library profession’s response to this problem is to advise libraries to apply their judgment and ingenuity in identifying appropriate peers for comparison purposes.

In the case of library rating systems—such as the German Library Association’s BIX ratings, HAPLR, and the new *LJ* Index—this problem is magnified significantly. Rating systems assign libraries to peer groups based on simplistic and imprecise indicators such as community population or library expenditures. Beyond ignoring possibly significant imprecision in these data, this creation of equivalent classes also ignores key differences on factors such as community demographics and needs, library mission, institutional context, and others. As a result, accuracy and validity of final rankings from these systems are compromised.

### Higher Statistics Are Always Better

As already noted, there are no objective criteria for evaluating library statistical indicators. This lack was identified by Altman in her description of a classic public library performance study in the 1970’s:

The project team was philosophically opposed to the practice of standard comparisons [of libraries by means of enabling or use measures] because of the arbitrary way in which they were set and the general lack of care used in making the comparisons. Had we taken it upon ourselves to pronounce that certain numbers were “good” or “bad,” we, too, would have been rightly accused of being arbitrary. . . . The study team felt strongly. . . . that each library staff should decide for themselves whether the findings for that library were acceptable in terms of performance expectations.<sup>20</sup>

More than three decades later, we still rely on this inadequate solution to a difficult problem. Libraries are advised to avoid automatically interpreting higher statistical

indicators as reflections of better performance, and lower statistical indicators as signals of poor performance. Beyond this, libraries receive no further guidelines for drawing final conclusions from statistical comparisons. As a result, they have only their own ingenuity to apply to this task.

Library rating systems, however, are exempt from any obligation to interpret comparative statistics judiciously. Instead, the algorithms used by these systems assume that higher statistical data unequivocally indicate better performance.<sup>21</sup> Without this assumption, comparative rating scores could not be calculated at all. Yet, this methodological compromise weakens the meaningfulness of library ratings as measures of performance.

### All Loans Are Not Equal; All Visits Are Not Equivalent

Holt and Elliot (2003) maintain that “All circulations [loans] are not equal” and “All visitations do not represent equal consumption of services or equal value to the library customer.”<sup>22</sup> To this we can add that all materials of a given format (books, video recordings, electronic resources, and so on) also are not equal. Approaching this idea from the user perspective, Kyrillidou observes that “perceived quality as judged by the user does not relate to the extensiveness of resources or activities in a library.”<sup>23</sup> As already noted, extensiveness (counts) of materials and services communicate little about the relevance, quality, value, content, complexity, or other significant characteristics of library services and resources.

These observations are direct challenges to the legitimacy of the traditional equivalences upon which public library statistics have been based for more than a century. Nevertheless, public library national rating schemes add, subtract, divide, and otherwise combine these numbers without regard for the *homogenized* (to use Desrosières’ term) nature of the data. Both individual comparisons and aggregate comparisons, the latter in the form of library ratings, overlook key details of the actual phenomena that library statistics represent. For this reason library comparisons based on standard statistical data are quite limited in meaning.

Beyond statistical definitions that homogenize data, traditional library statistics do not tap more sophisticated dimensions such as library mission, collection quality, match between services and community needs, contribution to community quality of life, and so on—characteristics necessary to make judgments about library performance, merit, excellence, and value. Again, this lessens the significance of library statistical comparisons and summary rating systems based on these.

### Interpreting Measured Constructs

A crucial step in utilizing library statistics is the interpretation of the ultimate meaning of the measures (as opposed the meanings of magnitudes that measures might take on). In this task libraries are again left to their own devices to decipher what the measures might

mean. Certainly, libraries can avail themselves of measurement approaches prescribed in library assessment literature, such as assessments of service quality, quality management, and performance scorecards. However, these approaches provide limited guidance in drawing inferences from library statistical measures. Nevertheless, drawing these inferences is a crucial step in the overall assessment process. Each step in the process—from conceptualizing assessment questions, designing measurement tools, collecting and analyzing data, to formulating conclusions—needs to be performed carefully to assure the quality of assessment results.

Let us consider how this inference generally occurs in a typical library assessment effort. As an example, I suggest one library indicator from the global statistics model developed by IFLA, the UNESCO Institute of Statistics, and ISO—*seats per capita*. Interpreted literally, this indicator is a measure of the amount of physical seating capacity in a library divided by population counts. More abstractly, the indicator can be seen as a reflection of a library's commitment to promoting accessibility to information resources. It could also be evidence that a library fosters in-house utilization of materials, accommodates disabled or elderly patrons, or strives to portray library buildings as comfortable and welcoming locations. Most abstractly, seating capacity may be viewed as a singular indicator among a larger set of indicators that, together, reflect a more generic library attribute one might call “overall performance.”

By means of this example, we see that, for each individual statistical indicator, a small set of possible interpretations can be derived (presuming that only reasonable inferences are to be considered). Further, these interpretations can vary from concrete to abstract. With library ratings, however, the process of associating individual indicators with various concepts is sidestepped. When library statistical data are combined into single summary scores, these scores are perceived as measures of a single, if somewhat vague, concept of *overall library performance*. Even if we were to qualify this perception by asserting that performance is multi-faceted (as reflected in the component statistics used in formulating ratings), the format of single-score ratings still implies that they measure a unitary attribute of libraries. It is tempting to describe this summative attribute using terms like *quality*, *excellence*, *effectiveness*, or *value*. However, as stated already, standard statistical data fail to tap key library dimensions necessary for representing these more sophisticated concepts. For this reason we should avoid using these terms in this context.

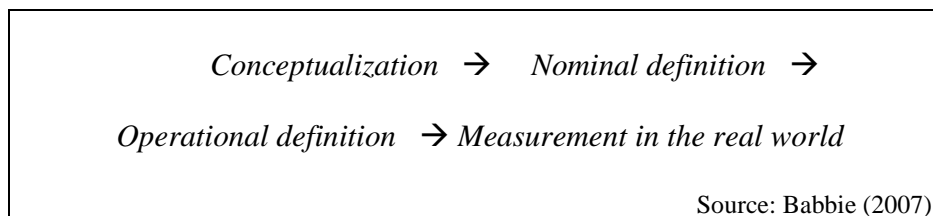
Library ratings exemplify the concept that Desrosières called *encoding*, mentioned in the introduction of this paper. The equivalence class is *overall library performance* and its definition is the formulas utilized by the ratings. These formulas utilize some, but not all, of the available library statistical indicators. When these selected indicators are summarized into a single score, again, certain details of interest are included while others are excluded.

This leaves us with the question of what library ratings actually mean? That is, what generic concepts are represented by the specific selections of library indicators used in public library rating systems? I suggest that the library assessment community can play an important role in guiding libraries in answering this question. And I believe that this

task can be facilitated by exploring measurement practices in social and behavioral science research. A basic understanding of these research protocols should help libraries understand the importance of interpreting data carefully and systematically. It should also provide libraries with a greater appreciation of the complexity of the measurement process, itself.

Babbie describes social science measurement as involving the sequence shown in Figure 1.<sup>24</sup> First, an initial research concept or *construct* is identified, for instance, *customer satisfaction*. Then, a nominal definition is established based upon consensus among professional expertise in the field being studied. From this, a more specific definition stating precisely what data will be gathered as measures of the concept. Next, after a measurement tool is developed and tested, it is used to obtain measurements from a real-world setting.

Figure 1



In social science research, the initial research construct is also referred to as a *latent variable*.<sup>25</sup> This designation signifies that the construct cannot be observed directly. Instead, only perceptible phenomena considered to be evidence of the existence of the unobservable construct can be measured. Due to both the complexity of the construct and its inherent unobservability, multiple indicators are usually required to assure that an adequate reflection of the underlying phenomenon is obtained.

Once data are collected and analyzed, researchers then draw inferences from the operational measures in order to make statements about underlying construct(s). Thus, this process requires moving between abstract constructs, their more intermediate meanings, and the operational (more concrete) measures. Patterns and relationships detected in and among these measures are proposed as reflections of patterns and relationships existing in and among the latent variables, that is, the concepts being studied (for example, customer satisfaction, service quality, user attitudes, and so on).

This type of measurement process is implied, but not explained, in popular approaches to performance measurement in library assessment and management literature, such as quality management, business excellence, and balanced scorecards. As a result, linkages between data collected and relevant performance concepts (constructs)—and vice versa—are not nearly as straightforward as these approaches suggest. Our task is to bring these methodological ideas to bear on library assessment practice so that conclusions drawn from performance data will be sound, justifiable, and appropriate.

## Recommendations

As the availability of national and international library statistics increases, use of the data for comparing libraries is inevitable. Library statisticians may well experiment with regional, national, or international library rating systems. In preparation for this possibility, the library assessment community should help libraries understand the problems inherent to these comparative exercises. Since we lack adequate methods for identifying peer libraries, and standard data collection is characterized by both homogenization and inevitable imprecision of data, libraries need to be advised of the importance of interpreting ratings results cautiously. The same is true for comparisons of individual libraries with each other.

Given the significant methodological limitations of aggregate library ratings, one might ask why these ratings should be designed and published at all? I suggest that the ratings can be useful for showcasing libraries and drawing attention to the need for further information about library value and effectiveness. Library ratings are best viewed as *contests* rather than as rigorous measurement exercises. With contests, it is quite legitimate to accept arbitrary restrictions as conditions necessary for conducting the competition. As long as libraries and library stakeholders are educated about these methodological compromises, then ratings scores can be recognized as simplistic, broad-brush feedback about library performance. While some libraries may take pride in scoring very well in these limited ratings, the exercise may inspire all libraries to pursue more sophisticated evaluation measures to describe and help improve their performance.

## Notes

1. Alain Desrosières. *The Politics of Large Numbers: A History of Statistical Reasoning* (Cambridge, MA: Harvard University Press, 1998).
2. Thomas Hennen. "Go Ahead Name Them: America's Best Public Libraries," *American Libraries* 30, no. 1 (1999): 72-76.
3. Ray Lyons. "Unsettling Scores: An Evaluation of the Hennen Annual Public Library Ratings," *Public Library Quarterly* 26, no. 3/4 (2007): 49-100.
4. Keith Curry Lance and Ray Lyons. "The New LJ Index," *Library Journal* 133, no. 11 (2008): 38-41.
5. Thomas Childers and Nancy A. Van House. "The Grail of Goodness: The Effective Public Library," *Library Journal* 114, no. 16 (1989): 44-49; Nancy A. Van House, Beth T. Weill, and Charles R. McClure, *Library Performance: A Practical Approach* (Chicago: American Library Association, 1990); Philip Calvert and Rowena Cullen. "Performance Measurement in New Zealand Public Libraries: A Research Project," *APLIS*, 5, no. 11 (1992): 3-12.
6. Roswitha Poll and Peter te Boekhorst. *Measuring Quality: Performance Measurement in Libraries*, 2nd ed. (Munich: KF Saur, 2007); Suzan Imholz and Jennifer Weil Arns, *Worth Their Weight: An Assessment of the Evolving Field of Library Valuation*, Americans for Libraries Council, 2007. <http://www.actforlibraries.org/pdf/WorthTheirWeight.pdf> (Accessed July 1, 2008).
7. Peter Hernon and Ellen Altman. *Assessing Service Quality: Satisfying the Expectations of Library Customers* (Chicago: American Library Association, 1998), 9.
8. Another rationale for collecting statistical data is to promote the value of libraries to key stakeholders. This use assumes that library statistical data are suitable for evaluating library performance.
9. The United States Congress passed the Government Performance and Results Act of 1993 in order to establish results-oriented management in federal agencies. The U.S. government classifies performance measures into five types: count of products/services provided (outputs), measures of operational efficiency, measures of customer satisfaction, measures of product/service quality, and outcome measures. United States General Accounting Office. *Results-Oriented Government: GRPA Has Established a Solid Foundation for Achieving Greater Results* (Washington, D.C.: U.S. General Accounting Office, 2004, GAO-04-38), 12.

10. Nancy A. Van House et al. *Output Measures for Public Libraries: A Manual of Standardized Procedures*, 2<sup>nd</sup> ed. (Chicago: American Library Association, 1987), 1.
11. Rowena Cullen. "Does Performance Measurement Improve Organisational Effectiveness? A Postmodern Analysis," *Performance Measurement and Metrics* 1, no.1 (1999): 9-30.
12. Beryl A. Radin. *Challenging the Performance Movement: Accountability, Complexity, and Democratic Values* (Washington, DC: Georgetown University Press, 2006).
13. Dean R. Spitzer. *Transforming Performance Measurement: Rethinking the Way We Measure and Drive Organizational Success* (New York: American Management Association, 2007); Gloria A. Grizzle. "Performance Measurement and Dysfunction: The Dark Side of Quantifying Work," *Public Performance and Management Review* 25, no. 4 (2002): 363-369.
14. Elaine Morley, Scott P. Bryant, and Harry P. Hatry. *Comparative Performance Measurement* (Washington, DC: Urban Institute, 2001); David N. Ammons. *Municipal Benchmarks: Assessing Local Performance and Establishing Community Standards* (Thousand Oaks, California: Sage, 2001).
15. David N. Ammons. *Municipal Benchmarks*, 7.
16. Roswitha Poll. "Benchmarking with Quality Indicators: National Projects," *Performance Measurement and Metrics* 8, no.1 (2007): 41-53; Ignace Glorieux, Toon Kuppens, and Dieter Vandbroeck. "Mind the Gap: Societal Limits to Public Library Effectiveness," *Library and Information Science Research* 29 (2007): 188-208; Joseph R. Matthews. *The Evaluation and Measurement of Library Services* (Westport, Connecticut: Libraries Unlimited, 2007).
17. Roswitha Poll and Peter te Boekhorst. *Measuring Quality*.
18. David N. Ammons. *Municipal Benchmarks*.
19. Elaine Morley et al., *Comparative Performance Measurement*, 6.
20. Ellen Altman. "Reflections on Performance Measures Fifteen Years Later," In *Library Performance, Accountability, and Responsiveness: Essays in honor of Ernest R. DeProspero*, C.C. Curran and F.W. Summers, eds. (Norwood, NJ: Ablex, 1990): 13.
21. In the case of measures of operational efficiency, lower values would be considered indicators of better performance.
22. Glen Holt and Donald Eliot. "Measuring Outcomes: Applying Cost-benefit Analysis to Mid-sized and Smaller Public Libraries," *Library Trends* 51, no 3 (2003): 425.

23. Martha Kyrillidou. "From Input and Output Measures to Quality and Outcome Measures, or, From the User in the Life of the Library to the Library in the Life of the User," *Journal of Academic Librarianship* 26, no. 1/2 (2003): 44.

24. Earl Babbie, *The Practice of Social Research*, 11<sup>th</sup> ed. (Beaumont, California: Thomson, 2007).

25. Robert F. DeVellis. *Scale Development: Theory and Applications*. 2<sup>nd</sup> ed. (Thousand Oaks, California: Sage, 2003), 14.