

Community College Alumni as Customers: A discussion of the  
benefits and risks of a common student data repository

W. Thomas Hamlin  
Extensible Information Systems (EIS)  
Lansing, MI  
(517) 887-3087  
[thomas@thomashamlin.com](mailto:thomas@thomashamlin.com)  
[thomas.hamlin@davenport.edu](mailto:thomas.hamlin@davenport.edu)

Prepared for:  
Michigan Association for Institutional Research  
16<sup>th</sup> Annual Conference  
October 16-18, 2002  
Lansing, Michigan

Copyright © 2002 by W. Thomas Hamlin

All rights reserved under International and Pan-American Copyright Conventions.

No part of this paper may be reproduced or transmitted in any form or by any means electronic or mechanical including but not exclusive to photocopying, recording, or by any information storage and retrieval system without the expressed written permission from the author or his estate.

For permission contact: Document Manager  
EIS  
Lansing, MI 48910-5631  
info@thomashamlin.com  
(517) 887-3087

# Abstract

This white paper provides an overview of the present state of community college evaluation processes and procedures, discusses possible enhancements, and outlines the basic design concepts for a unified system to collect, organize and manage student data. The intended audience is community college administrators, researchers and others who frequently utilize data/information from their existing systems and/or who may be searching for alternatives to existing evaluation processes and procedures.

Higher education's evaluation and review teams, accreditation agencies, policy makers and governing boards have in common interest in assessing college outcomes. Education finance experts want to know if expenditures and investments are worthwhile. College administrators sometimes need ready answers to questions about how students feel about aspects of their college experience. (McLure and Valiga 2002)

Four general topics will be discussed. The first topic is the overview of the limitations of existing evaluation processes and procedures. The second topic is the importance of data quality to any successful community college operation. The third topic is the utility of external data and the fourth topic is a high-level system design including an outline of a Student Input Process (SIP) and the differences between On-Line Transaction Processing (OLTP) and On-Line Analytical Processing (OLAP) systems

## Present State of Evaluations

The State of Michigan is blessed with 28 of the finest Community Colleges in the country. The staff and faculty are dedicated to educational excellence. The communities they serve, the students, and the student's employers are enriched by their efforts. In the face of this resounding success comes the requirement to better inform all stakeholders of the college's continued success in an increasing competitive world. Many of the evaluation processes and procedures offer a limited and restricted view of a college's real success since standardized evaluations do not account for the variability between students in their life situations and lifestyle.

For example the standard for gauging timely progress toward a two-year degree according to the National Graduation Rate Survey (GRS) is the 150 percent rule. A student is expected to complete an Associates degree program within 36 months from date of admission to a community college. Recent research indicates that two-year and four-year institutions vary considerably in their student's preparedness and goal-centeredness and therefore in the students ability and desire to complete within a pre-determined time frame. (Floyd, 2002)

Sometimes the evaluation criteria may be expressed as numerical indicators of (manufacturing production statistical process control (SPC)) quality that may be unfamiliar to the college administrator. Therefore misunderstandings may occur about the criteria, their meaning and application. (Jordan, 2002) For instance, an establishment of a 'baseline' may be an arbitrary single value taken from a past year where as 'baseline' in a SPC context is the average of many values over an extended period of time.

In addition, each of the 28 community colleges is responsible for the analysis, development, implementation and maintenance of their unique student data repositories. Because of the decentralized nature of Michigan's community college environment, there is little incentive to formally exchange student information between institutions.

Another issue is that standardization of data across independent organizations is often overlooked when repositories are designed and constructed. Further, many important data elements are entirely overlooked in the development of data repositories, which severely limit research flexibility. There may be potentially several systems at each college containing duplicate or redundant student data. All in all, this makes it very difficult to perform cross-institutional longitudinal evaluation studies without considerable additional effort. In conclusion, there is an urgent need to standardize student data within and across all community colleges in order to obtain a balanced view of student outcomes and college evaluations.

The following sections define and expand upon the important concepts and ideas relating to a unified student data repository.

## Data Quality

While no data repository is entirely error free, errors can be minimized to an acceptable level. (English, 1999)

There are many challenges in the area of data quality to overcome. One notable example was a hospital emergency room (ER) admissions application. Hospital administrators, regulators and third party payers noticed that many ER admissions were initially coded as 'broken arm'. However, the release diagnosis could vary widely. Upon researching the situation, the real reason was discovered. 'Broken arm' was frequently used because the ER admissions application required an initial diagnosis. This presented an impossible situation for the admissions clerk. In most cases the admissions clerk found it easier to use 'broken arm' as a default entry rather than wait for a release diagnosis before processing the admission. Unfortunately, the application could not be changed, and data quality suffered as a result.

A similar situation could exist at a number of community colleges across the state. It is not uncommon for a variety of systems to be independently built without regard to downstream data, expanding informational needs, or coordination with other existing systems.

### Data Format

An example of an inappropriate data format that may be found at community colleges is a student's contact telephone number. Telephone numbers may be presented in a wide variety of formats in printed materials that are unsuitable for a database. In this example, a human being has little difficulty resolving all the vague and misleading queues represented in the different formats for phone numbers, however, 'our assistant', the computer cannot resolve these variations. Below are telephone formats commonly found in operational systems and in printed form. Refer to Table 1.

Table 1

<b>Common Telephone Number Formats</b>
9895551212
(989) 555-1212
989 555 1212
989/555/1212
989.555.1212

Each example has an entirely different resolution in the computer's memory. Moreover the number, 9895551212 could represent the population of microbes in a petri dish or angels dancing on the head of a pin. Without the proper documentation to describe and define the data in question, 9895551212 is just a number out of an infinite universe of numbers. The number has no meaning in isolation.

## Data Inconsistencies

Data inconsistencies can offer another challenge to data quality. Inconsistencies can derive from simple duplication even within an individual college's family of applications. Best's Law states that if data resides in two places, *it will be inconsistent*. If two different systems collect and maintain contact phone number information and the numbers are different, which is correct? Is this an error or simply an individual giving a home phone number at one time and a cell phone number at another? Variations and updates to area codes offer another prime example of the possibility for inconsistent data.

## Data with a Single Meaning

From another perspective, sound database design principles require the use of scalar/atomic values within database cells. Scalar values are singular in meaning. For example, nine is a scalar value. 9NCDUS is a code (not scalar) for the Ninth North Carolina District in the United States. As redistricting occurs one can only imagine the problems related to coordinating the code changes in an unknown number of down stream systems. The number of down stream systems is unknown because we can never be sure what systems have adopted the code. Only when these down stream systems fail does the breadth of the problem become apparent.

However, scalar values can represent complex ideas. For example, 989 could represent a telephone area code. As such, it should be the only value found in cells in the column labeled TelephoneAreaCode. The same holds true for 555, the Telephone Exchange Number, and for 1212 the Telephone Number.

## Data Flexibility

An additional benefit to a well-designed data solution is flexibility. With the recent number of new Telephone Areas Codes within Michigan, a flexible design would be greatly appreciated by those responsible for maintaining telephone data. Systems based on sound design principles are less expensive to maintain.

## Data Standards

Data that meets accepted standards could be crossed checked and updated. For example student contract information could be verified against the United States Postal Service (USPS) change of address information, the telephone company to verify active working phone numbers, and other third party providers to enhance its quality and usefulness. If the data does not meet a common standard then each time the data is crossed checked and updated many complex manipulations of the data must occur for the process to be successful. This increases the time and expense of such operations.

## Data Validation Approaches

A significant benefit of data standards for a proposed unified student repository system is that student data can be validated with top down (research-oriented processes) and

bottom up (direct student participation) efforts. This two-pronged endeavor ensures the best quality data available short of ‘perfect information’ or a government mandated system.

The top down approach is research focused. In other words, the data and information gained from formal surveys or other research methods are expensive and time consuming and therefore they are done infrequently. When they are conducted they offer an important opportunity to validate a student’s information. The opportunity that is too valuable to miss. Any modifications to a student’s information should be updated in a unified database.

An innovation offered in this proposed system is the inclusion of data and information supplied and validated directly by the student (bottom up). Students would be interested in sharing and accurately maintaining their personal information (bottom up) if it is within the limits of ‘privacy fair use’ and they gain something (a consideration) in return. A consideration could be Internet access, a community calendar, discount tickets, easier access to common administrative functions, or other innovative programs that are offered by a community college.

### Describing the Data (Metadata)

Metadata is simply data about data. Metadata greatly expands the usefulness of the data to those who may not be familiar with it. An example will serve to outline the benefits. All computer systems store data as a series of binary numbers, that is, 1’s and 0’s. It is literally impossible for humans to interpret the series of digits, such as 010110110 as anything meaningful without Metadata. Modern computer database management systems allow for the use of Human readable text to describe objects like tables, columns, stored procedures and triggers. A database table could be named Customer Contact Data, a column could be named Customer Contact Telephone Exchange Number, a stored procedure could be named Update Data Dictionary with Column Name Change and finally a trigger could be named Upon Update Insert New Data Into Customer Contact Data.

Metadata can provide even more information to the system user. It is often a requirement to provide a Data Dictionary to fully describe the meaning of the data. For example, Customer Contact Telephone Exchange Number would be more meaningful if the description said, “Customer Contact Telephone Exchange Number is the telephone exchange number of the local residence of a student when attending classes at our community college”. Returning to this example, the data description could be simplified for the end user’s use to “Customer Contact Telephone Exchange Number is the local telephone exchange number of a student”.

Make your Metadata useable by providing easily accessible and understandable documentation. Metadata repositories and Data Dictionaries are like home exercise equipment, the best one is the one that gets used.

A word of caution is in order. Whole books have been written outlining in precise detail how to create Metadata. While these tomes are of benefit to database design academics

and theorists, they can be too burdensome for the practical application of a working system. It is beneficial to the end user of a system if data descriptions (Metadata) utilize our everyday word usage. More detailed and precise data definitions should be reserved for the dusty documentation on the shelf.

### The Value of Perfect Information

The ultimate goal of a system is to provide the best information possible to the end user. End users in this case could be researchers, administrative assistants, college administrators, marketers or others. The cost/benefit ratio of data needs to be examined. When examined carefully, the cost of 'perfect information' is prohibitive. To obtain 'perfect information' on the total population of community college students, past, present and future, one person would have to live their life simply to document another's every move. Truly, this is not a viable option. Therefore, the cost of information must be related to its value to the organization. Some data and information is critical, some is important and some is 'nice to know' but not worth the effort to capture, organize and maintain.

## Utility of External Data

We can only imagine the kinds of data and information that could be correlated with student data. However,

Regardless of the benefits of recent advances in methodology, there are still hurdles that must become with gathering data for multi-institutional research. (Cohoon, 2002)

It is not the usefulness of external data (if it exists) that prohibits its use. It is the inability of the existing systems to incorporate external data and information.

### Data Availability from External Sources

Despite all the problems, community colleges as a group are expected to provide data for many purposes. There are few viable options by which a community college can provide data about the progress of their students after they leave. One option is to conduct surveys. A recent survey was conducted using a complete listing of each institution's students in the form of self-adhesive mailing labels. The survey provider randomly removed a predetermined number of labels from the total provided and contacted the selected individual by phone. One can only imagine the improvements that could be made to the process with an adequately designed, fully featured data repository.

Another option is to gain access to external government data stores. These repositories offer great potential because they are mandated by legislation and they already exist. It would seem to be a simple thing to obtain access to this data since it is already 'in house'. In a recent effort, wage and hour information from Michigan's Unemployment Insurance (UI) data repository was analyzed as a potential source of data. This data repository tracks an individual's earnings and employers over time. Limited approval was gained to

use the data, but because of privacy and security concerns, 'third parties' cannot have access to the data. The restriction of third parties limits the resources available for data analysis and processing. Each community college would be required to individually 'recreate the wheel' to utilize the data. Also, there are significant gaps in the data. Federal employees, the self-employed or those persons who have moved out of Michigan are not included in the UI data. While the data is still valuable, the value is limited.

### Multiple Uses for Student Data

Community colleges have many uses for student data. They need to promote new programs and course offerings to students and to follow up with other important information. They need to conduct research projects. They need to follow up with employers about student outcomes. However, the community college face a unique situation: few groups have a common interest in following up with former community college students over a long period of time. Further, Alumni associations are non-existent or very rare at community colleges. Simply put, the same data repository can be used to meet a variety of data and informational needs if it is available.

The larger an issue-oriented consortium is, the more likely it can deliver the following advantages: a richer data context for comparative analyses, a more diverse pool of institutions for constructing meaningful comparison groups and a lower cost for individual members. (Smith, 2002)

### Data Privacy and Ownership

Gathering data from a variety of sources and moving it across networks create a number of concerns for the community college, two of which is student Privacy and Confidentiality.

Data privacy at its core is a question of ownership. The only acceptable answer in our democratic society is that the individual ultimately owns all their personal data. Confusion about ownership begins when individuals trade the 'privacy fair use' of their data for 'a consideration'. A consideration in this context means that individuals are willing to give up some privacy in order to gain something of value. In exchange for a consideration the data owner gives an organization permission to use their data if processed correctly. Examples of considerations are car loans, admissions to a community college, government benefits or quality medical care.

There are strong differences of opinion however. Many groups in our society view government data repositories with a skeptical eye. References to George Orwell's 1984 are all too common. Some persons and groups frame data privacy issues in a religious context using the Biblical books of Daniel, Matthew and Revelation as their guide. Regardless of the frame of reference, those who abuse another's personal data do so at their own risk. They also put the community college at risk.

To summarize privacy for this white paper, the assumption is that individuals retain the ownership of their data at all times. As long as privacy is not abused, there is little need

for concern on the part of a researcher or marketer in using student data for well-defined purposes.

## **System Design**

This section discusses the high-level design of a proposed system that could collect, organize and maintain community college student data for a variety of purposes. It includes a discussion of design independence, an innovative Student Input Process (SIP) process, the two basic systems types, users, system security and architecture.

### **Interaction of System Design and The Implementing Technologies**

The system design should be vendor/technology neutral to the greatest extent possible. All technology vendors, software or hardware, are proprietary or limiting to a certain degree. The best method to minimize this problem of technological dead ends is to make the design of the system as independent of the implementing technology as possible. The most direct path to independence is for conceptual and logical models/interfaces to be designed without regard to a specific vendor or technology (i.e. Microsoft vs. IBM, Java vs. Windows, COM vs. COBRA, Oracle vs. SQL Server 2000, etc.). Techniques based on abstract concepts such as personas, goals and scenarios lead to durable designs that any vendor can implement. (Cooper, 1999).

### **The Student Input Process (SIP)**

An important innovation of this proposed system is the development of a SIP to assist the community college in the maintenance and validation of a student's data and information. This is referred to as the bottom up process in the Data Validation Approaches section above.

The community college understands that personal information deteriorates over time. Individuals marry, divorce, move, use cell phones exclusively and more. To provide valued added services to the individual and to provide quality information to the community college, it is paramount that information over a long term be kept up to date and as be error free as possible.

The incentive for the student to maintain their data could be to preserve access to the desirable features of a college's Web based information system (Portal) or other attractive features. The basis for a SIP might be annual data verification of a student's current and contact information, employment status, income and other outcome variables.

The access to desirable features is a consideration of the type discussed in the Data Privacy and Ownership section above. In addition, individuals may desire to keep up with news or new services offered by their community college. The valued added services in a portal that could be available to the existing and former students are found in the table below. Refer to Table 2.

Table 2

<b>Suggested Valued Added Services</b>
Continuing education credits seminars
Community Calendar
Transcript requests
Employment opportunities
Additional course offerings (credit and non-credit)
On campus events
Community service opportunities
Organizational Portals
And more

To restate the issue, in the normal course of events community colleges are fully committed to education but do not have the resources to maintain student contact, employment or other important information consistently over time. Only periodically will a survey capture a small sampling of this data. Therefore, the college loses track of the success of past students and frequently does not have a 'single unified view' of their present students. Further, other stakeholders, i.e., government, businesses, community organizations, or taxpayers are left to their own devices to determine the success of a community college. It is preferred to be proactive with information about a community college rather than reactive.

The proposed system could therefore provide data for a variety of purposes and maintain a high level of data quality with combination validation processes.

### On-Line Transaction Processing versus On-Line Analytical Processing Systems

This section discusses the basic difference between On-Line Transaction Processing (OLTP) systems and On-Line Analytical Processing (OLAP) systems. Essentially OLAP systems are reporting only systems while OLTP systems process the business data for an organization.

OLTP systems are the lifeblood of the organization and the main focus of the Information Technology (IT) or Management Information Systems (MIS) departments. Examples of OLTP systems are Student Enrollment, Accounting, Human Resources and Inventory systems. OLTP systems most commonly provide a series of standard reports on the data and information within its boundaries.

Most of these systems were developed independently of each other for a particular use within a department such as Accounting. Almost without exception, if the system needed a piece of information, i.e. a student's name, it would provide a utility for capturing the data and maintaining the data in its computer code base. Rarely does the format of the data resemble a similar piece of data in another system. It is even more rare to find independent systems that use an external source of data for its internal purposes. There are numerous reasons for this situation but the results are the same, duplicate data and

information are manifest across the organization. This condition is called information stove piping.

To solve this dilemma, many organizations are implementing Enterprise Resource Planning (ERP) systems that provide one unified system for all common functions across the organization. Examples are SAP, Oracle, Great Plains Business Solutions and PeopleSoft. There are others.

Several points are important here. First, to implement one of these ERP systems, the organization must be willing to accept the design and process flow of data and information provided by the software vendor. Unique features of existing custom applications are difficult at best (if not impossible) to duplicate within the ERP system. Without question, any customization of an ERP system is expensive. The benefit of ERP systems is that they provide consistent data across all the installed modules.

Second, operational reporting stemming from an ERP system focuses primarily on '*What has happened?*' This question is in the past tense and provides little understanding of the external factors affecting the future of the organization. More analytical type questions cannot be answered with static operational reports.

Third, every organization has unique requirements that can only be satisfied by custom applications (there are many). It is best to leave these applications running independently and not attempt to integrate them into an ERP system. This leaves important stovepipe applications running that need additional maintenance. Further, they are not integrated with other systems for reporting purposes.

This mixed environment presents several additional critical problems. The first critical problem is that since the organization can only answer the operational '*What has happened?*' question with a standard ERP report. It has a very limited ability to answer the analytical question of '*Why did it happen?*' And certainly ERP systems are not designed to answer the most important analytical question of '*What will happen?*' OLAP systems were created to fill this gap. They offer deeper analytics for a variety of purposes.

For instance, imagine if your organization could not project an increase in operational usage (therefore operational expense) for the coming period. The reader will immediately recognize the need to take operational data and manipulate it for analytical purposes. Other readers understand that Lotus 1-2-3, QuattroPro and Excel are spreadsheet tools made expressly to fill this need to analyze operational data.

A second critical problem is the availability, accessibility and usefulness of operational data. For example, the finance department in every organization has an annual budget and forecasting process that uses spreadsheet tools to analyze their data. Typically each department contributes several dozen numbers to the process. Re-keying this small amount of data, while not error free, is certainly possible. Can you image trying to re-key several thousand (millions or billions) cells of student data? If possible at all, the process would be filled with human input errors.

To overcome this situation, OLAP systems get their data primarily from the organization’s operational stovepipe OLTP and ERP systems through an automated extraction, transformation and loading process (ETL). Typical OLAP systems can have dozens of operational data sources. This manual/automated capability gap is an opportunity for an integrated OLAP reporting system.

The third critical problem is the temptation for an organization to blend the operational and OLAP systems into a single system. This is tantamount to a flying boat. It might work for small applications, but this system architecture is extremely limited in larger applications.

To briefly summarize this section, operational systems exist to support business functions and provide static ‘*What has happened?*’ reports. OLAP systems are used for variety of purposes and to provide a utility for answering additional the two analytic questions, ‘*Why did it happen?*’ and ‘*What will happen?*’ The two systems should forever remain separate and distinct. Refer to Table 3.

Table 3

<b>System Type</b>	<b>Function</b>	<b>Reports</b>	<b>Questions</b>
OLTP/ERP	Operational	Static	What has happened?
OLAP	Decision support, marketing, research	Flexible and analytical	Why did it happen? What will happen?

**Potential Users of Data**

The initial research indicates that there are four potential users (persona) of the data in a Student Data OLAP system. The first persona is the student. In exchange for current and accurate information about themselves, they could use the SIP process to maintain access to the desired features of a portal system.

The second persona is the professional researcher who would access the data to research questions of interest. This data would be the basis for answering the analytical questions mentioned above. The system would be designed so that the student data would remain under a community college’s control. However, anonymous demographic information would be available to other ‘member researchers’. This research persona could be augmented to provide information to a third college administration persona in the form of a ‘digital dashboard’. A digital dashboard could provide summary information such as current enrollment by department by semester, student retention and satisfaction or other important metrics. Everything an administrator would need would be on a single screen.

The fourth user is the community college marketer/surveyor/administrative assistant persona who would use the system as a front end to facilitate their efforts. An example of features that could be available for surveys is the ‘automated presentation’ of a random sample from a desired population. Further, the surveyor could update the student contact data as part of the validation process.

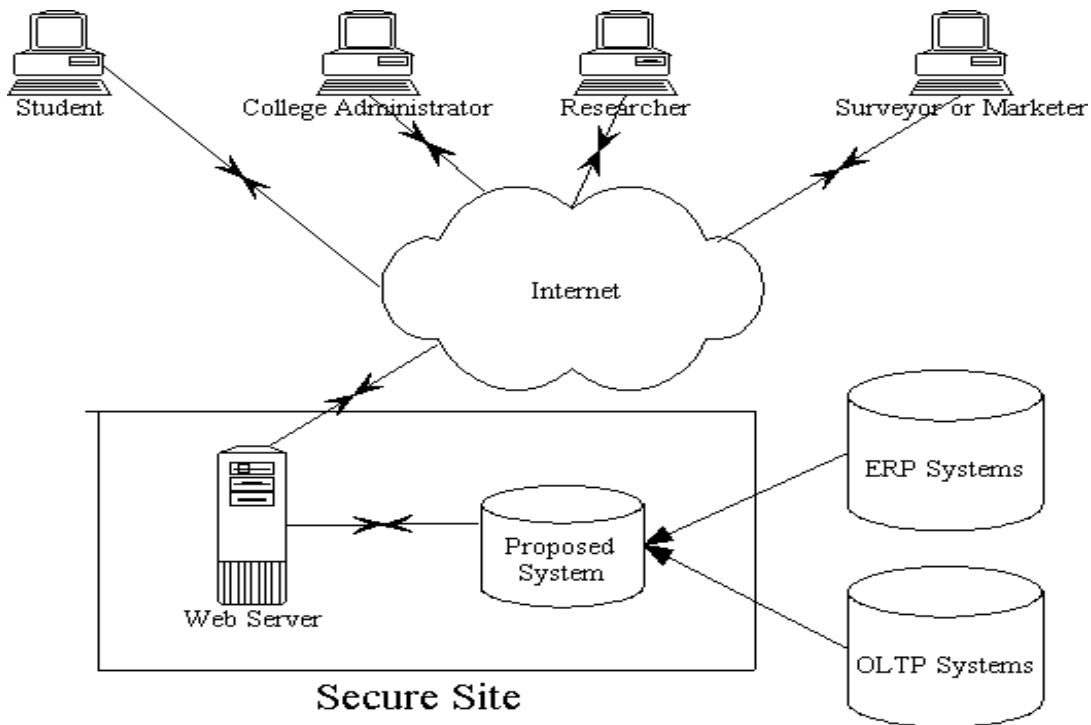
### Physical System Security and Privacy

The system would have several features that would enhance physical security and privacy of the student data. The first feature would be sound security policies and procedures. The second feature would be the separation of student identifying data (SID) from student demographic data. The third feature would be additional sign-in security to access identifying student information. It would consist of a domain name, user name, password and biometrics. The biometrics could consist of a fingerprint-identifying device attached to the user’s computer. The fourth feature would be a set of security enabling technologies, such as encryption, Secure Socket Layers (SSL) and certificates. The final security feature could be physical enclosure of the computers themselves.

### High Level System Architecture

Figure 1 is a high-level system architectural diagram of the proposed system. Several features are obvious. The first feature is that access to the system is through a Web Server. The second feature is that all the servers (therefore the data) are in a secure location. The third feature is that a separate server performs the processor intensive computing. This makes the system highly available through browser technologies, such as Internet Explorer or Netscape, and therefore less expensive to implement while supporting the required security. Refer to Figure 1.

Figure 1



## Summary

This white paper has attempted to advance the discussion about the usefulness of a standardized repository for collection, organization and maintenance of student data with in the secondary educational environment. Unique features outlined are the discussion of personal data ownership, partnering with the student to maintain their data over time and the common interest by the community colleges in developing a student data repository. The question is not whether a student data repository should be implemented but when.

## Bibliography

Cohoon, J. McGrath. 2002. Data for Multi-Institutional Research (431). *Association for Institutional Research 2002 Annual Forum Toronto, Canada June 2-5, 2002.*

Cooper, Alan. 1999. *The Inmates Are Running the Asylum. SAMS, A Division of Macmillan Computer Publishing, Indianapolis, Indiana.* ISBN: 0-672-31649-8

English, Larry P. 1999. *Improving Data Warehouse and Business Information Quality. Wiley Computer Publishing, John Wiley & Sons, New York, New York.* ISBN: 0-471-25383-9

Floyd, Nancy D. 2002. So How Long Have You Been Here? Using Retrospective Transcript Data to Examine Time to Completion at a Community College (616). *Association for Institutional Research 2002 Annual Forum Toronto, Canada June 2-5, 2002.*

Jordan, Larry. 2002. Accountability Indicators from the Viewpoint of Statistical Method (610). *Association for Institutional Research 2002 Annual Forum Toronto, Canada June 2-5, 2002.*

McLure, Gail T. and Valiga, Michael J. 2002. The Factor Structure Underlying Perceived College Outcomes (133). *Association for Institutional Research 2002 Annual Forum Toronto, Canada June 2-5, 2002.*

Smith, Teresa Y. 2002. The Consortium for Student Retention Data Exchange: A Case Study of Consortium Development (432). *Association for Institutional Research 2002 Annual Forum Toronto, Canada June 2-5, 2002.*