



Center for
Educator Compensation
Reform

Data Quality Essentials

*Guide to Implementation:
Resources for Applied Practice*

Jeffery G. Watson
University of Wisconsin–Madison

Sara B. Kraemer
University of Wisconsin–Madison

Christopher A. Thorn
University of Wisconsin–Madison

Data Quality Essentials

In order to complete compensation reform successfully, many school systems must transform information systems that were originally designed for reporting and accountability into systems that support performance-pay work. However, using data systems in new ways can quickly expose previously unnoticed data quality problems. The goal of this article is to help school systems identify, address, and plan for data quality problems before performance decisions are put under the scrutiny of system stakeholders.

This module focuses on the data quality challenges that states, districts, and schools must resolve when they reform compensation systems to take into account performance measures such as student achievement, teacher evaluation, and professional development. To begin, the following key questions must be addressed:

1. What are the key characteristics of quality data for compensation reform projects?
2. On what data quality problems should TIF project leaders focus?
3. What are some ways in which data quality problems can manifest within a compensation reform project, and what are some potential solutions to those problems?

The Teacher Incentive Fund (TIF) has awarded more than \$80 million to 34 local and state education agencies to support the design and implementation of performance-pay systems. In addition to TIF grantees, many other school systems are also examining and implementing performance-pay systems. These projects all have significant information technology (IT) components because districts generally use measures drawn from many data sources to determine individual pay amounts, including assessments, student enrollment, human resources, and teacher and principal observations. Because most districts use these extensive systems to make a relatively small number of decisions, there is the inherent tendency for what would be otherwise isolated data quality issues to be magnified within the performance plans.



It is imperative that TIF leaders know which teachers are teaching which students.

Figure 1. Data-centric process model for TIF projects.

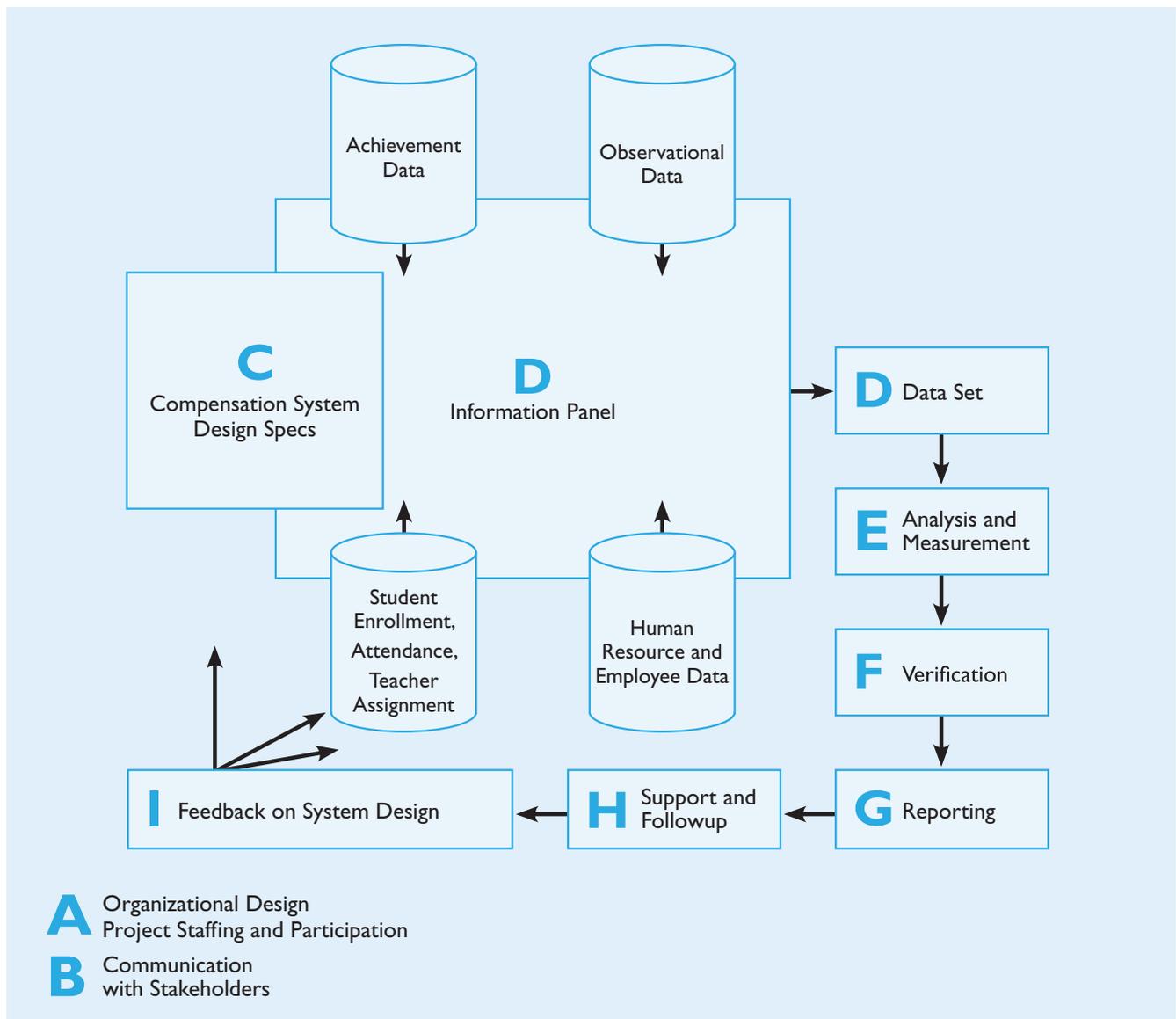


Figure 1 represents a process view of a TIF project implementation and takes a data-centric view of program implementation. Elements A and B represent the organizational and communication background within which all of the process elements take place. All of these systems and process elements (C-I) require staffing and are a part of the internal operations of the district. Block C (Compensation System Design Specs) represents the design specifications for the compensation plan. Design specifications drive the requirements (D) for human resources and student demographic, test, and observational data.

Those data are then validated against both existing data cleaning rules and against new data quality requirements imposed by the compensation model (E). Indeed, we find that the implementation of new analytical requirements (particularly student-teacher links, by subject) often leads to another round of data validation and verification (sometimes including major changes in requirements for the operational systems from which these data were drawn).

When TIF leaders have validated their analytical results as sufficiently robust (an iteration between E and F), they can then report the results of the

compensation model (G). This reporting is supported with follow-up outreach, training, and additional supporting materials (H) to help participants in the system understand the reports and make use of the results to reflect on their school and personal performance. Finally, TIF projects should apply discrepancies in the reporting, feedback from stakeholders, and data gaps back into the system (I) through refinements in the compensation model and in improvements in data collection and manipulation in the source systems from which the data were drawn.

This data-centric process model differs significantly from how most districts utilize their data systems. Most districts that base teacher pay on years of experience and highest degree earned use their information systems within a fairly constrained scope. Namely, districts generally use their student information systems (SIS) to enroll students in schools, schedule students and teachers into course sections, and track attendance and possibly disciplinary actions. In addition, districts traditionally rely on human resource (HR) systems to track employee data and deliver payroll and on their assessment systems to meet accountability requirements at the school and grade levels. The traditional uses of HR and SIS data systems are often insulated from each other and have little overlap between organizational, technical, or workflow processes. However, school systems implementing compensation reform must use all of these systems to make a single systemic decision: how much to pay teachers and principals. Data quality problems from any one system have the potential to affect a compensation reform project in negative ways.

The goals of this module are to identify the dimensions of data quality from a compensation reform perspective, provide TIF project leaders with a data quality focus, and describe common data quality problems and solutions. To address the first goal, this module presents six dimensions of data quality that are key to understanding how information systems must evolve to meet the needs of TIF projects.¹

Data quality problems from any one system have the potential to affect a compensation reform project in negative ways.

Individually, these dimensions represent design requirements for overhauling district data systems for the purpose of building decision support capacity. Based on work with several large U.S. districts across multiple projects, these dimensions focus attention on the functional role of data and information systems within decisionmaking. These dimensions do not specify a data model per se, nor do they specify content (e.g., prescribe a data dictionary). The dimensions complement the work by the Schools Interoperability Framework Association (SIFA)² and the Data Quality Campaign by focusing attention on the role of data within the context of decision-making (i.e., determining performance awards) and the technology environments of large districts.

The second goal of this module is to help TIF leaders focus their attention on the quality of student-teacher linkage data. It is imperative that TIF leaders know which teachers are teaching which students. This section describes and defines student-teacher linkage data and presents common ways in which this kind of data can be corrupted. The third goal of this module is to present methods of assessing the quality of data as well as ways to improve data quality. This section provides real-world examples and solutions that have been tested in a large urban district.

What characteristics does a data system need to have to support a performance-based pay system?

Watson described six dimensions of data quality: accuracy, validity, granularity, interoperability, relational, and reducibility.³ The definitions of these dimensions are presented below (see Table 1).

Table 1. Dimensions of data quality

Data quality dimensions	Six dimensions of data quality
Accuracy	The degree to which data reflect reality.
Validity	The degree to which data measure an intended construct.
Granularity	The number of individuals (e.g., students), items (e.g., test questions), or period of time (e.g., semester versus yearly attendance) over which data are aggregated.
Interoperability	The degree to which data are integrated across data systems.
Relational	The degree to which an information system's underlying model of a data system is capable of capturing reality.
Reducibility	The degree to which data support the formation of categories of entities.

1. Accuracy

Accuracy is the degree to which data reflect reality. Are the data correct? This is a fundamental aspect of data quality and is probably one that easily comes to mind for most people when they confront the issue of data quality. It is not uncommon to see relatively large error rates in self-reported data. If one collected race and ethnicity data from students and parents through multiple avenues, it is normal to see correlations below .90 on those data. Racial group identification is a socially complex phenomenon and can be influenced by the reason for collecting the data. For example, on first entry to schooling in an open enrollment system, there may be an incentive for a student to identify with a particular racial category (for multi-racial students) if this will increase the likelihood of getting into a particular school or program. This might not be the racial or ethnic category normally chosen if the decision did not have a benefit associated with it.

Another common cause of accuracy problems is the lack of linkages (e.g., data that two or more systems can access and update) between student scheduling systems and HR systems. During the student assignment period in the summer, it is often possible to insert dummy teacher names as placeholders for teachers that the district plans to hire in the fall. These placeholder names remain in the system after the start of school because there is no formal link between HR and scheduling systems that would connect “New Teacher 1” with the school’s new math teacher *Bob Smith*. Other common causes of inaccurate data include poorly designed computer interfaces that do not check the validity of data at entry, inadequate training, and human error.⁴

2. Validity

Validity is the degree to which data measure an intended construct. Do the data, regardless of their accuracy, represent the attribute or variable that they are supposed to represent? In a simple example, the U.S. Census Bureau changed the way in which race and ethnicity are reported. Instead of using one variable to report both race and ethnicity (e.g., White, Black, Hispanic), the Bureau now reports race and ethnicity separately, so that it is now possible to differentiate race and ethnicity independently (e.g., Black Hispanic versus White Hispanic).

One common example in education is the student school of record. While most students do not change schools during an academic year, many do, especially in urban settings. Thus, the school at which students are tested may not be the school at which they received most of their instruction. Because school-level student achievement measures become increasingly invalid as the number of mobile students increases, many districts will hold schools accountable only for those students who were enrolled for a full academic year. In this case, student achievement measures for a given school lose validity as the percentage of mobile students increases.

3. Granularity

Granularity is the number of individuals (e.g., students), items (e.g., test questions), or period of time (e.g., semester versus yearly attendance) over which data are aggregated. Data quality suffers when the data granularity does not support the analytic lens, or unit of analysis, of decisionmakers.⁵ For example, in urban districts, student mobility is often cited as a problem for schools because students who are mobile are exposed to different forms of instruction and different curricula across a single year as they move between classrooms. Attempting to control for the amount of time a mobile student spends between two schools requires student-school data to be sampled frequently. However, many districts capture student-school linkages between one and three times per year (usually for determining budgets). Under-sampling student-school linkages limits the fidelity to which student learning can be attributed to schools or teachers. Some districts have implemented periodic or monthly verifications to begin to address the scope of this problem.

4. Interoperability

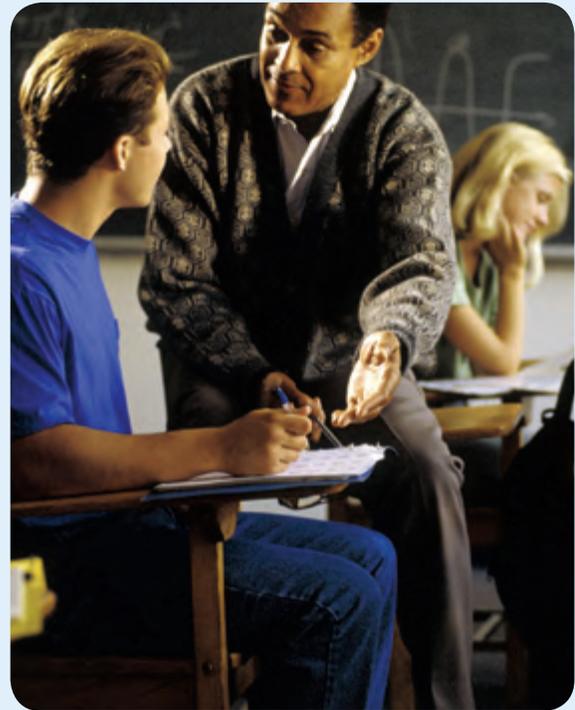
Interoperability is the degree to which data are shared across data systems. Generally, information systems in school districts are not integrated, although the SIFA has made significant progress toward establishing a unifying data model for developers. However, most school systems do not currently have a high degree of interoperability between source systems. An example of poor interoperability comes from a large urban district that recently attempted to merge teacher certification data from its HR system with teacher course assignment data from its SIS. In theory, records for the teachers in these two systems should have matched, but in reality, only about 80 percent of the records matched on name or identification number.

There are many reasons why data quality usually suffers when systems are not integrated. For example, when systems are not interoperable, data

migration is cumbersome. Thus, if an SIS is not integrated with the HR system, staff must enter teacher data twice, which increases the likelihood of spelling and typographical errors. In addition, if teachers change their last names when they get married, staff must update teacher data in two systems, rather than just one. This requires staff to match records between the two systems using a combination of automated and manual methods, a process that is likely to be both expensive and difficult.

5. Relational

Relational is the degree to which an information system's underlying model of a data system is capable of capturing the complex details of day-to-day schooling. When a data model is not able to capture the reality of a school, there is little hope that the data system will provide data that reflect what really



Student achievement measures for a given school lose validity as the percentage of mobile students increases.

happened within that school. For example, many SISs do not capture alternative approaches to course scheduling. Most systems allow schools to enter one teacher assignment for each course. When teachers decide to team-teach, or otherwise collaborate during instruction, it becomes difficult, if not impossible, to record teacher assignments accurately. Other examples of scheduling approaches that are difficult to capture from an SIS include block scheduling, remediation interventions (e.g., pull-out instruction, tutoring), and special education instruction.

6. Reducibility

Reducibility is the degree to which data support the formation of categories of entities. For example, teachers are often labeled as math or science teachers or as a teacher of a particular grade. Categorizing teachers as either math or science teachers when they actually teach across content areas would be an over-reduction of teacher assignment data. Likewise, assigning one school code to students who are mobile is an over-reduction of student enrollment data. Many times the causes of over-reduction of data lie in how data are pulled from source systems and pushed into a repository (e.g., a data warehouse). That is to say, the over-reduction of data (e.g., excluding mobile students' alternate schools) sometimes occurs after data have been extracted from the SIS.

These dimensions should be used to foster dialogue among program directors, district policymakers, and IT staff. Ideally, administrators will engage IT staff in early discussions about these data quality dimensions for all sources of data that will be used to determine performance awards. Only staff that are intimately familiar with the systems that collect and manage the relevant data will be able to accurately assess many of these aspects of data quality. Without this kind of collaboration, projects risk incorrectly awarding bonuses. The result of such a misstep could be serious and provide perverse incentives around less productive forms of teaching and school organization.

We propose several types of questions to consider when applying the data quality dimensions to performance-based pay systems. Table 2 presents a set of questions to give district staff a sense of how they might begin conversations with IT staff.

Table 2. Using data quality dimensions to guide discussions between project leaders and IT staff

Data quality dimensions	Sample questions to ask IT staff
Accuracy	Are student-teacher linkages in the student information system (SIS) correct? Do teacher records in the SIS match teacher records in the human resources (HR) system?
Granularity	Do data support using a unit of analysis that matches the performance-pay systems (e.g., individual teacher bonuses)?
Validity	Are performance metrics consistent with other performance measures? Do student-teacher links captured in the SIS reflect those in classrooms?
Interoperability	Can students be connected to teachers and other instructional staff? How much work will be involved in making sure that individuals (e.g., students and teachers) match across systems?
Relational	Is the SIS data model able to capture secondary student-teacher linkages?
Reducibility	Are teachers of multiple subjects incorrectly identified as teachers of only one particular content area? Do categories represent all teachers?

Focusing on Data Quality: Student-Teacher Linkages

TIF leaders need to ensure that they are making decisions about teacher and principal compensation based on data that are of sufficient quality. This section provides TIF leaders with a road map of factors that can affect the quality of the data. Drawn from the experiences and challenges encountered by previous grantees, this road map will illuminate where and how to look for possible problems.

Student-teacher linkage data describe which teachers are teaching which students. Without high-quality student-teacher data, TIF programs would struggle to make even the most basic progress toward compensation reform. Luckily, most districts have systems and staff in place to deliver these critical data. However, the quality of these data should be carefully assessed

before any awards are made. There are two reasons for assessing the data quality of student-teacher linkages as early as possible. First, this information is critical for computing valid performance awards. Second, fixing data quality problems can be time consuming and difficult, and the older the data the more difficult and time consuming this work becomes.

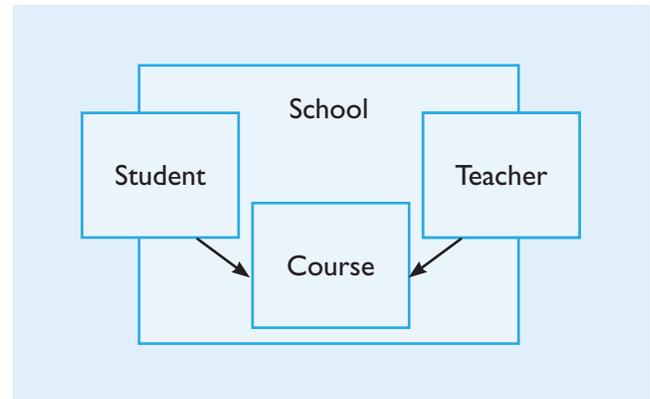
For all but the smallest of districts, data about student-teacher linkages originate and are maintained in an SIS. Several points are worth making about SISs. First, they are transactional in nature. This means that the data within the system are ever changing. An SIS should be thought of as a snapshot of the various scheduling, attendance, and discipline transactions that occur within schools at a specific point in time. Since the data within an SIS are always changing, data at one point in time will be different from data at a later point in time.

Second, these systems have a large user base and are affected by social and technical factors. A large number of staff who vary in technical expertise, experience, and organizational role contribute to the operation and management of an SIS. An SIS user base includes data entry personnel, teachers, counselors, school administrators, district technical staff, and sometimes contractors and consultants. SIS data are also affected by workflows and processes that are distributed over long periods of time. For example, scheduling usually begins eight to nine months prior to the beginning of the school year. SIS data can also be affected by software and hardware updates. Given the complexities of the typical SIS, it is highly recommended that TIF projects include technical experts at every step of their project. This might mean that departments that don't traditionally work together find themselves working very closely together.

Third, most SISs do not link students directly to teachers. Instead, the SIS will associate a student with a school when the student is admitted or registered and later with a course when the student is enrolled into one or more courses. Through

completely separate processes, teachers are also connected to individual schools when they are assigned to one or more sites. Teachers are later assigned to specific courses prior to the school year and possibly reassigned after student enrollment. Thus, student-teacher linkages are more accurately portrayed as a student-course-teacher-school linkage (Figure 2).

Figure 2. SIS Student-Teacher linkages. Students are associated with schools and courses through admission and enrollment processes. Teachers are associated with schools and courses through separate processes.



Thus, multiple processes mediate student-teacher linkages. In addition, these processes vary considerably within districts and even schools. When contemplating data quality of student-teacher linkages, TIF leaders should be concerned with being able to answer four basic questions:

1. What is a student?
2. What is a teacher?
3. What is a course?
4. What is a school?

Based on CECR work with TIF grantees, TIF projects may experience some difficulty when answering these questions. Variation within districts and schools creates nuances that an SIS cannot capture. Sometimes the SIS data model cannot capture the reality of school schedules and instructional models. Thus, the data in an SIS will not always map onto reality. The probability of a mismatch between SIS data and reality increases with district size and the degree to which schools within the district are differentiated.

What is a student?

This module defines *students* as the learners enrolled within an educational entity. The question that most TIF projects have to consider is when to include a student within the dataset(s) used to calculate awards. Non-typical students, for example, mobile students (students who move between classrooms and schools), present interesting decisions for TIF leaders. Since mobile students are often disproportionately associated with schools in neighborhoods with lower average income, excluding mobile students may undermine the spirit of many TIF programs aimed at improving student achievement at high-needs schools. Likewise, schools that are purposefully designed to serve special populations of students (e.g., adjudicated students, teen mothers) may also present challenges to TIF programs in urban districts. Districts often use non-traditional schools to serve students with extraordinary circumstances (e.g., teen mothers, adjudicated students). These schools in turn often use organizational and educational strategies that challenge the traditional student role that is found in main stream schools. TIF leaders should consider whether or not any schools within the district are designed to function in a way that challenges the validity of the TIF award structure. Low-attending students present a more mundane example of how the definition of students can affect TIF compensation systems. High absence rates weaken the degree to which student learning (or lack of learning) can be attributed to a specific teacher. Schools or teachers who are purposefully selected to teach and intervene low-attending student groups may be less likely to qualify for a performance award.

What is a teacher?

This module defines teachers as persons employed by an educational entity to educate students. TIF projects must assess the ability of data systems to track not just the traditional teacher, but also those education professionals serving in nontraditional roles. The latter includes long-term substitute



Excluding mobile students may undermine the spirit of many TIF programs aimed at improving student achievement at high-needs schools.

teachers, itinerant teachers, teachers who have been reassigned mid-year, and those who teach across schools or only part-time. Most districts will adjust the assignment of some teachers based on differences between expected and actual enrollment patterns. Sometimes teachers are purposefully assigned to be itinerant within a district. In addition, many circumstances can lead some teachers to miss more instructional time than others. Teacher leaders will often receive more professional development. Many teachers may take extended leave, and some teachers may be chronically absent from work. Some teachers teach across grades and content areas, and others are assigned to teach outside of their preferred content area.

What is a school?

This module defines *school* as any organization that receives funding from the school district to educate students. Again, TIF leaders must decide early which schools are to be included within the compensation systems. Some schools, especially those that are purposefully designed to offer innovative programming or serve traditionally hard-to-serve populations, may create instructional and organizational structures

that the SIS cannot manage. Schools within schools, charter schools, and non-instrumentality schools create complex and changing school units that may be difficult to manage within the SIS.

What is a course?

This module defines course as an organization structure that includes at least one teacher and one student and a curriculum. Course data are critical to a TIF project because students and teachers are connected through SIS courses. In addition, TIF projects classify teachers into content areas based on the content areas of the courses taught by the teacher. Even in medium-sized districts, the SIS contains tens of thousands courses. Documentation of these courses may be sparse, depending on how actively the district manages the course catalog. As a result, TIF leaders may have to guess which courses fall into math and reading categories. In addition, innovative schools intentionally create nontraditional instructional environments that are difficult for many SISs to track. For example, project-based schools may avoid enrolling students into courses until after the end of the semester or year. Other schools may use team teaching methods that are difficult to manage within the SIS. These factors lead to data in the SIS that may not represent reality.

Planning ahead: What programs improve teaching and learning?

TIF grantees build their compensation reform efforts on the assumption that rewarding effective practice will lead to diffusion of effective practices. However, TIF projects often provide little support designed to help teachers and principals *know* how to improve. When assessing student-teacher linkages, TIF leaders should consider what improvement programs exist within their district because access to these projects may be a key support for improving practice. Connecting a TIF program to a district's improvement efforts provides support to teachers who want to improve their effectiveness. Therefore, TIF leaders may want to begin tracking the significant

improvement programs within their districts. When district leaders begin tracking interventions, they raise the visibility of those programs significantly. Since these programs may be *owned* by other district leaders, it is important to consider how to track and connect these programs to a TIF project. TIF leaders who want to track programming may consider the following questions to guide information gathering.

- What are the content and focus of the program?
- What grade student population is most likely to benefit?
- What professional development is associated with the program?
- How does this program improve teaching and learning?

Common Causes for Poor Student-Teacher Linkages:

TIF projects require high-quality data, especially data that are related to connecting teachers to students. The following list contains common causes for low-quality student-teacher linkage data:

- Student mobility – decreases the amount of learning that can be attributed to a given teacher.
- Low student attendance – decreases the amount of learning that can be attributed to a given teacher.
- Teacher mobility – decreases the amount of learning that can be attributed to a given teacher.
- Low teacher attendance – decreases the amount of learning that can be attributed to the teacher.
- Inaccurate teacher data in SIS – makes identification of teacher difficult.
- Unrecorded instructional supports – weakens the validity of attributing student learning to the assigned classroom teacher.

- Nontraditional instructional methods – may lead to invalid SIS course enrollment data.
- Unique organizational models – may add instructional staff and programming other than those recorded in an SIS.
- Nonintegrated HR and SIS systems – increases the opportunity for data entry errors that prevent the connection of teachers in an SIS to teachers in HR or payroll systems.
- Poorly managed course catalogues/offerings – increases the difficulty of knowing about content of courses.
- Inadequate data entry staff job design, technical support, and training – leads to poor data entry practices, shortcuts, and competing pressures on data entry staff.
- Confusing or poorly designed interface – increases the opportunity for SIS operator error.
- Inadequate SIS data model – encourages schools to use work arounds in an SIS.
- Lack of data quality management – increases the rate of data errors within an SIS for a given school site.

Prior to meeting with project and technology staff, TIF leaders should define student-teacher linkages in a way that is measurable and easy to communicate. Basically, TIF projects need to know which teachers are teaching which students. However, most TIF projects do not try to measure performance for all teachers or all students. Should band instructors be included? How about special education resource rooms? TIF leaders should reference their project design to ensure that requests for student-teacher linkage data are consistent with the design of the overall project.

Assessing and Improving Data Quality

When school district staff members encounter data quality problems, they may try to minimize the risks associated with these existing problems, but to do so risks losing stakeholder support for the project. If data quality problems arise after awards are paid out, reactions will likely be very negative. We strongly caution school systems implementing compensation reform to anticipate and plan for data quality problems that may arise. Solutions are usually within reach, though project staff and IT staff will need to support corrective actions jointly.

Most likely, data quality problems will have both social and technological roots.⁶ For example, it might be tempting to blame inaccurate data on sloppy data processing, but it is important to assess whether the interface design of a data system causes an increase in data entry errors. More training may be needed for technology staff, or work pressures may encourage users to take shortcuts or otherwise subvert the system from its intended use. Regardless of the causes, school systems should consider prioritizing data quality issues to help guide efforts to improve data quality. Usually long- and short-term solutions will be identified.

The remainder of this module presents three examples from actual school districts to illustrate how these dimensions can be used to identify, assess, and improve data quality.

Anticipate and plan for data quality problems. Solutions are usually within reach, though project staff and IT staff will need to support corrective actions jointly.

Data Quality Challenge #1

Connecting teacher data from separate student information and human resource systems

Schools will have to merge data from the SIS with data from their payroll systems. As noted previously, integrating data across these systems is not always easy. In one large urban district, only about 80 percent of the teacher records in the district's HR system matched teacher records in the SIS. Some of the actual teacher 'names' that were entered in the SIS that could not be linked to actual teachers included:

Teacher A – MRP2	Teacher C – Sci6B
Teacher B – MRPI	Teacher D – Orchestra

In addition, some buildings used organizational structures that were not manageable with the SIS because the underlying data model did not allow teachers to assume multiple roles within the system. For example, team teaching was not easily managed by this district's SIS, and as a result, schools entered a workable schedule that did not accurately portray the real schedule used by the school.

Two types of analyses should be helpful for determining the extent to which inaccurate data compromise a district's ability to integrate data across systems. First, matching teachers in the SIS and the HR system reveals when data accuracy is lacking because inaccurate data result in incomplete matches. It may be helpful to summarize matches by grade and school. Second, understanding why inaccurate data are occurring involves analyzing the tasks and processes that might affect data quality. For example, in prior work with a large urban district, multiple factors led to poor data quality.

One glaring cause was that the two systems used two different identification systems, which required data processing staff to look up teachers' employee numbers and names in the HR and enter these data by hand into the SIS. If these systems were integrated, teacher identifiers would in most cases load into the SIS directly from the HR system. Another problem was found when data entry duties were analyzed. As is the case in most districts, schools followed a complex workflow that required data processing and administrative staff to create course catalogues and preliminary schedules before teaching assignments were made. Thus, staff sometimes had to create placeholders for teachers who would be hired in the future. Staff had to update teacher assignments once staffing was finalized (sometimes after the beginning of the school year). Failure to update the SIS correctly resulted in placeholder entries like 'Teacher A – MRP2.'

Potential Solutions to Data Quality Challenge #1

There are four potential solutions to these problems. They involve both social and technical interventions and both short- and long-term interventions.

1. Build data quality checks for data-entry screens that use heuristicsⁱ or look-up tables.ⁱⁱ For example, when an employee number is entered into a system, the system could check the number to make sure it conforms to the expected format (e.g., the correct number of digits). Better yet, data entry should be minimized whenever possible by pulling data from other systems rather than requiring the same information to be re-entered into a secondary system.

ⁱ Heuristic refers to using a problem-solving technique in which the most appropriate solution of several found by alternative methods is selected at successive stages of a program for use in the next step of the program.

ⁱⁱ Look-up tables refers to a strategy of storing data from one data system (e.g., HR data) in a table(s) so that they can be referenced by another data system (e.g., SIS). Look-up tables must be refreshed regularly.

2. Create data quality management tools (e.g., reports, training procedures) for district administrators to identify schools that need to improve data quality.
3. Build support for data entry staff (e.g., training, tech support).
4. Identify true needs of schools and develop use-cases in order to provide feedback to the SIS vendor and improve the underlying SIS data model (e.g., scheduling logistics).

Data Quality Challenge #2

Connecting Teachers to Students

Knowing which teachers taught which students is a critical linkage for school systems. However, SISs often do not support the complex organizational structures, such as team teaching and pullouts, that schools use. Moreover, the data model may not capture additional instructional support staff (e.g., pull-out specialists, instructors in after-school activities). Districts are not likely to capture all of the nuances of a student's school year. Instead, they tend to focus on identifying a teacher (and school) of record rather than on other instructors who also contribute to student learning.

In addition, SISs often do not record multiple roles of individuals or flexible organizational units. Ideally, an SIS should:

1. Link mobile students to multiple teachers.
2. Use course titles that reflect true curricular content.
3. Indicate team-teaching and link teachers to correct course content.
4. Link students to additional staff who provide instruction (e.g., in pull-outs, tutoring, and after-school programs), not just classroom teachers of record.



Once a student-teacher linkage file is assembled, it is important to verify these linkages with staff.

Assessing the accuracy and validity of the student-teacher linkages is a good first step toward knowing the extent to which this particular data quality issue presents challenges to a performance-pay system. One way to do this is as follows. First, identify where student-teacher linkages are easiest to track (e.g., elementary schools that use traditional organizational models). This simplifies the problem and makes analysis more manageable. Second, count the number of students assigned to each teacher and identify any outliers, such as teachers with too many or too few students. Third, examine these outliers more closely for patterns (e.g., some special education teachers may have taught a small number of students across multiple sites). Once a student-teacher linkage file is assembled, it is important to verify these linkages with staff. See the section titled Verification for more detail.

Potential Solutions to Data Quality Challenge #2

Three solutions to student-teacher linking problems were identified:

1. Build management tools, such as reports, that summarize student-teacher linkages (e.g., student counts by teacher, counts of teachers per building; identify which teachers teach across schools) and that can be used to target training and management solutions.
2. Examine the capacity of the SIS to track students' exposure to team-teaching, block scheduling, and pull-outs. Consider alternatives for collecting data from schools when these strategies are used. This solution helps assess the validity of the data.
3. Create incentives for schools to record the teacher of record accurately and verify this with teachers. For example, a district might require teachers to build a course roster from a list of enrolled students. Although this is redundant, it serves to validate the accuracy of the teacher-student links in the district's SIS. This solution also helps improve the validity of the data. Guilford County, North Carolina, for example, piloted a student-teacher linkage verification process in 2007 to provide opportunities for all teachers to review and confirm the names of each student that they taught. Although most student-teacher linkages were correct, teachers did catch some errors before the district performed the final analyses that would be used to determine performance awards. The process included three rounds of data verification to ensure that the linkages were accurate.⁷

Data Quality Challenge #3

Classifying teachers into categories

Middle school enrollment data from a large urban district provides an excellent example for this type of data quality challenge. The extent to which teachers instruct across grades and content areas can be assessed by a two-step process. First, student-level enrollment data need to be summarized by assigning course sections into content areas. This may require developing a case logic that identifies the content area of every course number.⁵ Second, counting the number of students per teacher grouped by grade and content area will reveal when teachers have students across grades and content areas. Analysis of the middle school enrollment data revealed that:

- 20 percent of middle school math and science teachers taught students within a single grade and single content area;
- 60 percent taught students across grades, but not across content areas;
- 10 percent taught students within a single grade, but in both math and science courses; and
- 10 percent taught across grades and across content areas.

If this district implemented a performance-pay system that rewarded individual teachers for work in core content areas, 80 percent of middle school teachers would be teaching more than one grade or content area and could be eligible for more than one award. This suggests that districts should carefully decide how they will determine awards for teachers of multiple subjects and grades and clearly communicate the eligibility rules to teachers.

Potential Solutions to Data Quality Challenge #3

Solutions in this example are a little less clear-cut because school systems often have to limit the number of awards that teachers can receive. One



Determine the number of teachers who teach multiple grades and content areas in order to project maximum costs of performance-based compensation systems accurately.

solution is to give a teacher an award for either 6th-grade math or 7th-grade math, but not both. Perhaps the easiest solution is to give a teacher a performance award for any grade or content area in which student performance meets specified criteria. Before doing so, however, the district should determine how this policy would affect the number of potential awards and program costs.

Another option would be to assign a teacher to the grade and content area in which he/she taught the most students. However, this might raise concerns about the compensation system, especially if teachers are assigned to high-need areas outside of their area of specialty. Although many human resource systems will note an area of expertise for teachers (e.g., secondary math), these data are often at odds with teacher assignment data in the SIS. We strongly caution districts that these data should not be treated as reliable until their accuracy has been verified. Regardless of the solution that is adopted, we encourage districts to determine the number of teachers who teach multiple grades and content areas in order to project maximum costs of performance-based compensation systems accurately.

Verification, Follow-up, and Feedback

Once a data set has been assembled, it is important to verify student-teacher linkages before making award calculations. In addition to verifying class rosters, TIF leaders may also want to confirm teachers' predominant content area and school assignment. The verification process should be designed to present teachers and principals with class rosters for all relevant courses. For each course, teachers should be asked to confirm that they were the primary instructor and that each of the listed students was in that course. The process should also allow teachers to explain or annotate when a wrong assignment has been made. For example, a student might have spent more time in a different classroom so that he/she could be grouped with other students of similar ability.

Once awards have been calculated, TIF leaders should consider how to provide follow-up and support to teachers who want to improve their effectiveness. TIF leaders should be prepared to provide technical assistance around the reporting mechanisms used by the project. In addition, TIF leaders should consider if follow-up outreach, training, and additional supporting materials will help teachers reflect on their effectiveness and *know* how to improve.

TIF leaders should plan to improve their overall data quality through continuous improvement methods by collecting feedback from participants and measuring key data quality indicators. For example, asking teachers and principals about the processes used to verify student-teacher linkages will guide improvements for subsequent years. Measuring the number of schools, teachers, and students with missing or incorrect information allows TIF leaders to design and implement strategies for improving data quality at the point of collection. TIF leaders should also solicit feedback from participants to improve how TIF analysis and awards are calculated and reported.

Summary

It is clear that each school system that wishes to put a pay-for-performance system in place has unique IT needs and varying capacity to retool systems to support such work. However, all school systems require high-quality data to implement an effective performance-based compensation system. Improving data quality requires understanding both social and technical roots, and efforts for improvement may be short- and long-term. The six data quality dimensions presented in this module can help school systems assess and improve their data quality, and we recommend that they begin having conversations about these dimensions with IT staff. Often, only those who work closely with an information system will know enough detail about how data are collected, stored, and organized to provide an accurate assessment of data quality challenges and effective solutions.

TIF projects should focus on student-teacher linkages since these data are a critical to calculating fair and accurate awards. TIF leaders need to understand

how their data systems collect and manage student-teacher linkage data and include technical experts in the design as well as implementation phases of their projects. Also, TIF leaders will want to talk with data system experts about the ways that student-teacher linkage data quality may be compromised.

TIF projects need to actively assess and measure data quality. Analyzing key features of data quality is the first step in assessing and improving data collection. TIF leaders will want to develop this work with technical experts to continue to monitor and improve TIF data quality.

As TIF leaders implement their systems, it becomes increasingly important to consider how their systems are designed to help participants know how to improve and become more effective employees. TIF leaders should also collect information that will inform how their systems can collect and analyze data more efficiently as well as how to improve reporting and dissemination of TIF data and award decisions.



The six data quality dimensions presented in this module can help school systems assess and improve their data quality, and we recommend that they begin having conversations about these dimensions with IT staff.

- ¹ These six data quality dimensions were first presented by Jeffery Watson, University of Wisconsin-Madison, at the National and International Workshop on School Information Systems and Data Based Decision Making. A copy of his conference manuscript, titled *Defining Data Quality for Decision Support Systems in Education*, can be found in the proceedings of that conference.
- ² **The Schools Interoperability Framework Association** (SIFA) is a data-sharing specification originally developed to allow information systems within K-12 districts to exchange data without requiring wholesale replacement of existing systems. It includes both clear definitions for core data elements, as well as secure methods for exchanging data. The Schools Interoperability Framework Association was founded to define the original standard and to provide a governance infrastructure for improving and expanding the standards-setting work. The SIF Association includes private software firms, state educational agencies, school districts, and higher education institutions. The association has also expanded to include international members. See <http://www.sifinfo.org> for more information on the standard, the association, and its members.
- ³ Watson, J.G. (2007). *Defining data quality for decision support systems in education*. Published in the proceedings of the ISMIS National and International Workshop on School Information Systems and Data Based Decision Making.
- ⁴ English, L. (2002). The essentials of information quality management. *DM Review*, 12(9), 34-44.
- ⁵ Thorn, C.A. (2001). Knowledge management for educational information systems: What is the state of the field? *Education Policy Analysis Archives*, 9(47). <http://epaa.asu.edu/epaa/v9n47/>
- ⁶ English, L. (2002). The essentials of information quality management. *DM Review*, 12(9), 34-44.
- ⁷ Guilford County Schools. (2007, October). Student-linkage verification. *Mission Possible Newsletter*, 1(2), 2. http://www.gcsnc.com/depts/mission_possible/pdf/October%202007%20Newsletter.pdf

Data Quality Essentials

Revised edition, August 2009

Jeffery G. Watson, University of Wisconsin–Madison

Sara B. Kraemer, University of Wisconsin–Madison

Christopher A. Thorn, University of Wisconsin–Madison

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the suggested citation is: Watson, J.G., Kraemer, S.B., and Thorn, C.A. *Data Quality Essentials*. Center for Educator Compensation Reform. U.S. Department of Education, Office of Elementary and Secondary Education, Washington, D.C., 2009.

The Center for Educator Compensation Reform (CECR) was awarded to Westat — in partnership with Learning Point Associates, Synergy Enterprises Inc., Vanderbilt University, and the University of Wisconsin — by the U.S. Department of Education in October 2006.

The primary purpose of CECR is to support Teacher Incentive Fund (TIF) grantees in their implementation efforts through provision of sustained technical assistance and development and dissemination of timely resources. CECR also is charged with raising national awareness of alternative and effective strategies for educator compensation through a newsletter, a Web-based clearinghouse, and other outreach activities.

This work was originally produced in whole or in part by the CECR with funds from the U.S. Department of Education under contract number ED-06-CO-0110. The content does not necessarily reflect the position or policy of CECR or the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by CECR or the federal government.



Center for
Educator Compensation
Reform

Allison Henderson, Director

Phone: 888-202-1513

E-mail: cecr@westat.com

