Validity Threats

Detection and Control Practices for State and Local Education Officials

Accountability Systems and Reporting (ASR) State Collaborative

Council of Chief State School Officers, Washington, DC
COUNCIL OF CHIEF STATE SCHOOL OFFICERS
The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

DIVISION OF STATE SERVICES AND TECHNICAL ASSISTANCE
The Division of State Services and Technical Assistance supports state education agencies in developing standards-based systems that enable all children to succeed. Initiatives of the division support improved methods for collecting, analyzing and using information for decision-making; development of assessment resources; creation of high-quality professional preparation and development programs; emphasis on instruction suited for diverse learners; and the removal of barriers to academic success.

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS
The State Collaborate on Assessment and Student Standards (SCASS) Project was created in 1991 to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. The Division of State Services and Technical Assistance of the Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects. SCASS projects accomplish a wide variety of tasks identified by each of the groups including examining the needs and issues surrounding the area(s) of focus, determining the products and goals of the project, developing assessment materials and professional development materials on assessment, summarizing current research, analyzing best practice, examining technical issues, and/or providing guidance on federal legislation.

Funding support for this paper was entirely from member states of the Accountability Systems and Reporting State Collaborative (ASR-SCASS). For information about ASR-SCASS and for 50-state Accountability Profiles, see the CCSSO website: http://www.ccsso.org/projects/Accountability_Systems/.

Council of Chief State School Officers
Valerie Woodruff (Delaware), President
Elizabeth Burmaster (Wisconsin), President-Elect
G. Thomas Houlihan, Executive Director

Julia Lara, Deputy Executive Director, Division of State Services and Technical Assistance
Rolf K. Blank, Director of Education Indicators

Copies of this report may be ordered for $10.00 per copy from:
Council of Chief State School Officers
Attn: Publications
One Massachusetts Ave., NW, Suite 700
Washington, DC 20001
202-336-7016
Fax: 202-408-8072
http://www.ccsso.org/publications/

Copyright © 2006 Council of Chief State School Officers
All rights reserved.
Validity Threats

Detection and Control Practices for State and Local Education Officials

Accountability Systems and Reporting (ASR) State Collaborative

Robin Taylor (DE), chair, Internal Validity Workgroup
Consultants: J.P. Beaudoin, Bill Auty, Pete Goldschmidt
ASR Coordinators: Rolf Blank, Andra Williams

Council of Chief State School Officers, Washington, DC
PREFACE

The passage of the No Child Left Behind Act of 2001 (NCLB) required all states to implement or revise accountability and reporting systems using a set of business rules prescribed within the statute. For most states, existing structures required significant redesign to address both state reform efforts and federal regulations.

In light of increasing resource demands, all state agencies have operational structures necessary to collect and process information. Many of these data are collected from the district and school level and are the foundation used in the production of both assessment and accountability results\(^1\). Numerous statewide testing programs, a major source of accountability data, have contractually-bound quality assurance plans to reduce error during the production of assessment results. Likewise, all centralized information technology systems use automated field specification audits to detect and reject information failing to meet predesign field criteria (e.g., numeric values in a student’s first name). Many times data dictionaries, end-user manuals, and extensive trainings are provided to local school districts to increase the accuracy and timeliness of data from the field. The information dependency of an assessment and accountability system has become significantly greater as student demographics and program memberships are directly linked to consequences and rewards within the systems. However, most production controls are cross-sectional quality management designs using single year events. Further, many accountability systems combine non-standardized, undocumented, and professional judgment techniques to examine score production, output distributions, consequence allocations, and unintentional consequences within the context of a state’s theory of action. Having moved past the point of simple NCLB compliance, state accountability systems have evolved to the point that it is necessary to develop more standardized techniques to improve the quality of accountability results.

The improvement process typically follows six generic steps: (1) examining current practices, (2) identifying priorities, (3) developing intervention plans, (4) monitoring implementation, (5) evaluating impact, and (6) revising procedures. Each step is dependent on the one before. This document will examine in depth the first two steps through the lenses of quality assurance, data integrity, end-user feedback, and credibility to the system and the state education agency. Further, it provides examples from state practices and suggests questions for state and local educational agencies to consider when constructing documentation and practices to fight validity threats to accountability systems.

---

\(^1\) Iowa and Nebraska do not operate state-level, centralized assessment and accountability systems; however, both agencies are dependent on aggregate results provided by local educational agencies.
# TABLE OF CONTENTS

PREFACE ........................................................................................................................................ ii

TABLE OF CONTENTS ............................................................................................................... iii

LIST OF TABLES AND FIGURES................................................................................................v

Section 1. Examining Current Control and Detection Practices................................................1
   1.0 Quality Assurance............................................................................................................1
   1.1 Pre-NCLB accountability...............................................................................................1
   1.2 Error sources: System inputs .......................................................................................2
   1.3 Error sources: Design logic...........................................................................................2
   1.4 Detection and control procedures ................................................................................3
   1.5 Examining current practices .......................................................................................4
   1.6 Identifying priorities ....................................................................................................7

Section 2. Improving Data Integrity using Validation Procedures..........................................9
   2.0 Data Quality Section.......................................................................................................11
   2.1 Guiding Question........................................................................................................11
      2.1.1 Documented procedures......................................................................................11
      2.1.2 Undocumented procedures................................................................................12
   2.2 Guiding Question........................................................................................................12
      2.2.1 Economically disadvantaged (Title I)...............................................................12
      2.2.2 English language learners (Title III).................................................................12
      2.2.3 Students with disabilities (IDEA)......................................................................13
   2.3 Guiding Question........................................................................................................13
      2.3.1 Student demographics, membership, and enrollment........................................13
      2.3.2 Assessment results ............................................................................................14
   2.4 Guiding Question........................................................................................................15
      2.4.1 Lesson learned ...................................................................................................15
      2.4.2 Caveats ..............................................................................................................15
Table of Contents (cont.)

Section 3. Using Feedback to Guide Training and Determine Impact ........................................16

3.0 End-User Feedback ...........................................................................................................16

3.1 Guiding Question ............................................................................................................16

3.1.1 Training format and focus ......................................................................................16

3.1.2 Targeted audiences ...............................................................................................17

3.2 Guiding Question ............................................................................................................18

3.2.1 Staff mobility ..........................................................................................................18

3.2.2 Sustainability .........................................................................................................18

3.3 Guiding Question ............................................................................................................19

3.3.1 Customized training ...............................................................................................19

3.3.2 Modular materials .................................................................................................19

3.3.3 Policy sanctions ....................................................................................................19

3.4 Guiding Question ............................................................................................................19

3.4.1 Lesson learned .......................................................................................................19

3.4.2 Caveats ..................................................................................................................20

Section 4. Promoting Quality through Mapping and Replicating Production .........................21

4.0 Credibility ......................................................................................................................21

4.1 Guiding Question ............................................................................................................22

4.1.1 Technical manuals .................................................................................................22

4.1.2 Standardized controls ...........................................................................................22

4.1.3 Professional screening and examination ...............................................................23

4.1.4 Programming reviews ..........................................................................................24

4.1.5 Non-standardized reviews ....................................................................................25

4.2 Guiding Question ............................................................................................................26

4.2.1 Time constraints ....................................................................................................26

4.2.2 Adequate staffing ..................................................................................................26

4.3 Guiding Question ............................................................................................................26
TABLE OF CONTENTS (cont.)

4.3.1 Comprehensive vs. Targeted .......................................................................................................................... 26
4.3.2 Third-party ...................................................................................................................................................... 27
4.3.3 Stakeholder reviews ...................................................................................................................................... 27

4.4 Guiding Question ............................................................................................................................................... 27
4.4.1 Lesson learned ............................................................................................................................................... 27
4.4.2 Caveats ............................................................................................................................................................ 28

Summary: Looking Back-Looking Forward ............................................................................................................... 29

REFERENCES ............................................................................................................................................................ 30

APPENDIX A: DSTP Quality Control Checklist ..................................................................................................... 32
APPENDIX B: Accountability Score Quality Control Procedures ............................................................................. 33
APPENDIX C: Quality Assurance Diagnostic Matrix (QADM) .................................................................................. 35
APPENDIX D: QADM Action Plan Template ........................................................................................................... 36
APPENDIX E: West Virginia Five Year Plan for Accountability Validity .............................................................. 37

LIST OF TABLES

Table 1. Selected Validity Threats, Detection, and Control Procedures ................................................................. 6
Table 2. QADM Decision Framework ....................................................................................................................... 8
Table 3. Effects of Recoding Students with Disabilities .......................................................................................... 10
Table 4. Sample Quality Procedures for the 2004 DSTP ....................................................................................... 11
Table 5. Accountability Score Quality Control Procedures ....................................................................................... 23

LIST OF FIGURES

Figure 1. Accountability Production Process ......................................................................................................... 2
Figure 2. Generic Process Control ............................................................................................................................ 4
Figure 3. Quality Assurance Diagnostic Matrix (QADM) ........................................................................................ 5
Figure 4. Decision Logic Flowchart-Nebraska .......................................................................................................... 24
Figure 5. Programming Flowchart-Accountability Decision Logic-OAI ................................................................. 25
Quality Assurance
Examining Current Control and Detection Practices

Local and state education officials promote data-driven decision making by implementing targeted production controls to ensure assessment and accountability data, judgments, and reports are accurate and credible. Activities such as data clean up are used to correct erroneous information before prescribed production dates; however, too often many data remain unchanged. This phenomenon occurs because either (a) the end-user fails to address the problem, (b) a solution set does not match the current business rules, and/or (c) the corrections are isolated and not implemented systemically. These types of events often have a direct impact on the magnitude and direction of nonrandom errors among accountability results. So the question arises, how much unwanted variance in the accountability score can be detected and controlled given the available resources?

Exploring this question requires examining three essential aspects of all accountability systems: (a) data integrity (inputs), (b) design logic and business rule implementation (production), and (c) score distribution and feedback (output). Data integrity is dependent on the collection processes used to solicit, obtain, and maintain information provided by schools and school districts. The design logic articulates how policy (now converted into a set of business rules) will combine information to produce accountability results. Score, label, and consequence results are further evaluated to discern unusual distribution patterns across varying units of analysis and time periods. Activities ranging from informal, random reviews to efficiency-driven quality assurance techniques are used to detect and control unwanted data anomalies.

Once results are produced, they must be accurately migrated into the reporting system. These quality assurance activities require an understanding of what organizational structures are in place, the degree of sophistication, and the competing resource demands before initiating a systematic improvement plan. The following section will begin exploring quality assurance practices associated with the production of accountability determinations.

1.0 Quality Assurance

The Quality Assurance section focuses on providing the conceptual foundation necessary to examine current quality assurance practices associated with error detection and control procedures. This section addresses:

- pre-NCLB accountability
- NCLB design logic
- error sources
- detection and control procedures

1.1 Pre-NCLB accountability

Accountability systems have become prevalent throughout the country as stakeholder groups seek information regarding school and district performance. The standards-based educational initiatives of the early 1990s were promoted during the reauthorization of the Elementary and Secondary Education Act of 1965 (ESEA), known as the Improving America’s Schools Act of 1994 (IASA). With the passage of IASA, many states, such as Kentucky, Louisiana, North Carolina, Texas, and others began developing accountability systems to augment their statewide assessment systems. Other states, such as Alaska, Arizona, Maine, and Oklahoma, addressed the federal requirements but continued to focus on state-level initiatives surrounding content standards and assessments. Further, Iowa and Nebraska maintained their unique local assessment and accountability processes along with capacity building structures to address state reform initiatives.
NCLB outlined an accountability system focused on assessing content standards using statewide assessment structures. Every state established baseline thresholds in English language arts and mathematics using methods outlined in federal regulations. For the first time, accountability systems were required to hold schools and districts accountable for the annual performance of subpopulations of students. Although subgroup performances were reported under previous Title I provisions, accountability consequences were not applied at this level.

1.2 Error sources: system inputs

The design put forth in NCLB aims to evaluate the consumption of educational services by demonstrating changes in targeted outputs (many in the form of annual measurable objectives (AMOs) for assessment results); however, the potential for ecological fallacy within this logic is significant. Based on the evidence required under NCLB, either the school in the aggregate did not reach its AMO (true positive) or data indicated otherwise (true negative).

Inputs from information management and assessment systems are essential elements of all accountability systems used throughout the country. Error detection and control techniques must reduce the magnitude of nonrandom data entering the production process. These techniques should be standardized, documented approaches within a greater quality assurance improvement plan guided by system needs, priorities, and available resources. Assumptions, rather than evidence, that nonrandom errors are being detected and controlled prior to entering the accountability production process threaten the inference that the entity under focus truly reached (or did not reach) their productivity goals.

1.3 Error sources: design logic

As outlined in Figure 1, the NCLB design logic combines legislated business rules and design specifications to evaluate subgroup, school, and district performance, defined in terms of Adequate Yearly Progress (AYP). This accountability design narrows the focus of school productivity, while increasing the chance of producing misclassifications. Hill and DePascale (2003) have determined that sampling error has a major influence on the reliability of misclassifying a school as making or not making AYP in accountability systems. The misclassification of schools that results in corrective actions (labels, school choice, supplemental services, etc.) poses a validity threat to the system. Efforts
to reduce the standard error and subsequently the misclassification of schools could inadvertently generate a new validity threat to the system. For example, if the system is attempting to measure the academic productivity of subgroups but most subgroup sizes fail to reach the minimum n-count requirement, the accountability decisions may not be valid. An accountability system in which school size and diversity influences the number of AYP judgments (and the probability of misclassifications) is operating incoherently; however, because the subgroup error terms are not independent, it is possible that one factor may dominate those identified as missing AYP. This phenomenon exists because many students have membership in one, two, or even three additional categories beyond the whole school and ethnicity subgroups.

Understanding the accountability system determinations is further clouded by the changes in the unit of analysis. Aggregating output indicators to the unit of analysis (e.g., school, district) across applicable grades provides a cross-sectional, time-bound summary of academic achievement. At a specific time, the comparative inferences between a school with 10 percent of its students scoring proficient or higher on the standards-based assessment and a school with 80 percent will be significantly different. These differences are eliminated once the unit of analysis changes to subgroups, as the 10 percent school has low performances across all subgroups, yet the 80 percent school may have only one. Because the aggregate of the parts are greater than one, it may be necessary to shift the paradigm being used to collect construct validity evidence. This shift suggests the existence of two constructs (subgroup and school productivity) that are not mutually exclusive, similar to the type of overlapping found in the relationship between reading and mathematic assessments.

1.4 Detection and control procedures

Most, if not all, states have quality assurance procedures in place for those data inputs used in their accountability system separate and apart from those used to ensure data quality in their assessment systems. (Many states have production process controls associated with their centralized assessment and information management systems along with an output review process, such as online verification of data and targeted districts to replicate results.) Some states, such as Nebraska and West Virginia have implemented quality management plans to examine the validity characteristics of their accountability systems. Other states, such as Arizona, Delaware, Ohio, Oregon, and others, have detailed their accountability business rules and production processes used to make AYP determinations.

Some states have implemented formal statistical controls within their decision logic. For example, Arizona uses a statistical process control in determining accountability appeals, while Louisiana has combined both quantitative and qualitative evaluations prior to making appeal recommendations to its state board of education. As rigorous as these controls are, few states have implemented the comprehensive, rigorous quality control processes found in the private sector, such as Deming’s Total Quality Management, Juran’s Company-wide Quality Management, and Six Sigma’s Define, Measure, Analyze, Improve, and Control (DMAIC) framework.

The private sector quality management guidelines outline how detection and control activities can occur within different production aspects, from input to production to products. For accountability, the basic production inputs are student-level data that are compiled and processed using the design logic outlined in state policies and procedures. Input data must meet specific field specification and business rule criteria prior to moving into the score production process. The score production process applies the design logic to those data meeting selected criteria; other data are disregarded or processed in other score production processes.

A concrete example of these processes is when student performance data cannot be assigned to any one building. In most states, the data are aggregated at the district and state levels rather than being disregarded. The score production process requires human resources individuals, including subcontracted elements, having the skills and resources necessary to complete the production tasks. Strategic capacity building efforts within the agency can ensure critical paths have adequate human
resources to prevent catastrophic failure. In other words, when critical production paths have single points of failure, the agency runs the inevitable risk of significant production delays. Figure 2 outlines a generic production process with key subcomponents outlining how inputs are manipulated during production, and then output quality is monitored. Manipulating this diagram within the educational context can help in identifying critical production paths and areas needing quality control activities.

![Figure 2. Generic Process Control](image)

**1.5 Examining current practices**

Educational entities and their agents must operate with full exposure to the general public. With some exceptions, taxpayer dollars are allocated, monitored, and regulated by oversight bodies. Accountability activities, part of the greater education system, must be able to ensure scarce resources are being used in the best interest of the public. Continuous improvement practices strive to maintain high quality through transparent practices along with credible decision making to promote consumer satisfaction with public services. The improvement process typically follows six generic steps: (1) examining current practices, (2) identifying priorities, (3) developing intervention plans, (4) monitoring implementation, (5) evaluating impact, and (6) revising procedures. Each step is dependent on the one before.

One approach to embarking on improvement processes is to document the entire accountability system in enough detail as to examine micro-processes. Because accountability systems are interdependent with other parts of the agency, typically information management and assessment, some consideration must be given to the interrelationships among these components. As accountability systems become more sophisticated, the production of technical manuals (DDOE, 2004; TEA, 2005) detailing the business rules and design logic will document the operational sequence used to produce accountability results.

Quality assurance procedures within accountability systems continue to undergo numerous changes as the overall demand for high quality data has increased exponentially since the passage of NCLB. Business rule complexity, short reporting timelines, and consequence assignments, combined with staff, budget, and time constraints, have made comprehensive error detection and control processes difficult for state and local officials. As a
starting point, the overall quality assurance practices must be examined using a macro-analytic approach. This approach must focus on organizing and examining quality control practices used in the production of accountability results for local and state systems. These accountability process controls would provide a foundation for more complex error reduction processes, especially those associated with consequential validity. Further, the interrelationship among accountability systems and their information management and assessment counterparts must be considered within the needs assessment.

To address this, the Quality Assurance Diagnostic Matrix (QADM) was developed using conceptual work found within the Systems Security Engineering Capability Maturity Model® (Carnegie-Mellon, 2003). The QADM comprises a set of quality assurance procedures for each of three educational systems: (a) information management, (b) assessment, and (c) accountability. These quality procedures comprise the horizontal axis of the matrix, while clustered within the primary operational area. In the vertical axis, a series of quality process characteristics are organized into five stages. The stages range from quality practices not being addressed to standardized procedures used to examine production efficiencies. The QADM requires the educational entity under examination to rate itself across the stage continuum for each system component. The resulting bar charted area (see Figure 3) displays a macro-analytical quality pattern across major systems within the agency. At this point, stakeholders are able to visualize quality assurance areas needing development and/or augmentation. To further assist in understanding validity threats within the identified areas, Table 1 (on the subsequent page) is a list of threats, detection procedures, and preventive measures used by some states.

![Quality Assurance Diagnostic Matrix](image)

**Figure 3. Quality Assurance Diagnostic Matrix (QADM)**
### Table 1. Selected Validity Threats, Detection, and Control Procedures

<table>
<thead>
<tr>
<th>Threat</th>
<th>Location</th>
<th>Detection</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluding students from accountability computation by manipulating authorized reporting codes</td>
<td>Information Management</td>
<td>Data audit exclusion codes of highly suspect and over-reporting of key exclusion codes</td>
<td>Require user authentication at school-level with district concurrence</td>
</tr>
<tr>
<td>Non-standardized assessment administration</td>
<td>Assessment</td>
<td>Item analysis examining atypical response patterns</td>
<td>Public reporting of irregularities using web-based interface or toll-free hotline</td>
</tr>
<tr>
<td>Dropout coding errors</td>
<td>Information Management</td>
<td>Selective and random auditing of exclusion codes</td>
<td>Automate local control procedures for targeted codes</td>
</tr>
<tr>
<td>Promoting special education students across high-stakes assessment grades via the IEP process</td>
<td>Information Management</td>
<td>Automated script detecting student promotion greater than one grade level</td>
<td>Provide error reports to local entities for validation purposes</td>
</tr>
<tr>
<td>Promoting high school students across the assessment grade</td>
<td>Assessment</td>
<td>Cross-reference assessment files with fall and spring enrollment to detect students enrolled, promoted, and not assessed</td>
<td>Establish policies requiring assessment participation for students legitimately promoted across assessment grade</td>
</tr>
<tr>
<td>New schools</td>
<td>Accountability</td>
<td>Districts requesting approval for new school operations, select and audit all new school data</td>
<td>Document new school audit findings within accountability manual</td>
</tr>
<tr>
<td>Reconfigured or reconstituted schools</td>
<td>Accountability</td>
<td>Districts requesting closing-opening schools, reconfiguring grade structures, and/or reconstituting schools resulting from redistricting (voluntary or required) flagged for score validation audit</td>
<td>Document all restructuring, recalculate baseline values, screen for business rule manipulation</td>
</tr>
<tr>
<td>Applying incorrect rating label</td>
<td>Accountability</td>
<td>Local agency replication using business rules detailed in technical manual. Programming error associated with implementing AND rather than OR within decision logic</td>
<td>Flowcharting programming expanded in conjunction with module replication within the agency</td>
</tr>
<tr>
<td>Inaccuracies in identifying students in denominators (including duplication)</td>
<td>Information Management System</td>
<td>Schools given opportunity via access to centralized information processing system to detect inaccuracies in preliminary student list developed with the agency for purposes of determining AYP participation rates. Conduct quality checks for each school, on enrollment count fluctuations for each of several uploads used throughout the year to identify potential inaccuracies in Full Academic Year (FAY) student status needed to determine proficiency rates</td>
<td>Provide schools with secure web access to the agency’s database during and after AYP appeals process. Allows for corrections by school officials by giving access to student records required in AYP computations. Spring semester pre-appeals check by schools on student lists and enrollment files.</td>
</tr>
<tr>
<td>Validity Threat</td>
<td>Location</td>
<td>Detection</td>
<td>Control</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Conflicting enrollment information</td>
<td>Information</td>
<td>Verify that districts are submitting all students for which the district is claiming state school</td>
<td>Automated comparison of enrollment data used for Average Daily Membership (ADM) compared to denominator for AYP, require local agency resolution</td>
</tr>
<tr>
<td></td>
<td>Management</td>
<td>funds</td>
<td></td>
</tr>
<tr>
<td>Identifying abnormal accountability</td>
<td>Accountability</td>
<td>Developing prior year/current year reports flagging abnormal increases or decreases in test</td>
<td>Produce audit reports posted for district to review and act upon prior to the production of final accountability report</td>
</tr>
<tr>
<td>data changes</td>
<td></td>
<td>administration events or student demographic information</td>
<td></td>
</tr>
<tr>
<td>Identify data element outliers</td>
<td>Information</td>
<td>Identifies outliers on targeted data elements (economically disadvantaged, FAY)</td>
<td>Notifies districts of their outlier status and requires data auditing, editing, or rationale</td>
</tr>
<tr>
<td></td>
<td>Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect enrollment on test records</td>
<td>Assessment</td>
<td>Validation report for districts within a secure website where test records are available. Preliminary aggregate record counts by school, grade, and test subject of the number of test records</td>
<td>Establish examination procedure and reassignment once districts have examined preliminary data. Display records from modified or unique administrative conditions</td>
</tr>
<tr>
<td>Inclusion of alternative schools</td>
<td>Accountability</td>
<td>Schools serving unique populations authorized by state and local regulations having received officials school approval</td>
<td>Generate automated report to monitor enrollment patterns within the school year, cross-reference with assessment</td>
</tr>
<tr>
<td>Demographics of alternative schools</td>
<td>Accountability</td>
<td>Alternative school populations are disproportionally represented by minority and/or special education students</td>
<td>Create enrollment by subgroup trend analysis using pre-NCLB and post or other benchmarks, while coordinating with onsite monitor by federal program staff</td>
</tr>
</tbody>
</table>

The validity threats listed in Table 1 are a small sample from a greater population. In most state and local agencies, quality enhancing activities are being used to detect, evaluate, and control nonrandom error; however, these activities tend to operate in a compartmentalized manner. This phenomenon is in part due to the maturity of different parts of the organization. The time and resource demands associated with production have created an environment that demands immediate results. Examining the risks inherent to the system is typically completed once production has started—results are already being produced and accountability judgments assigned. This was clearly the case with the implementation of NCLB’s assessment and accountability requirements. The law required implementation prior to a full examination of the safeguards necessary to ensure high quality results while promoting efficient resource allocations. Now three years into implementing NCLB, agencies are re-examining and adjusting their quality practices to improve credibility and attain greater efficiencies.

1.6 Identifying priorities

Every local and state agency should have a quality management improvement plan based on prioritized needs of the organization. In its simplest form, the action plan is a series of documented steps defining how a process should be conducted to achieve a standardized output amenable to validation reviews. A decision framework (Table 2) is outlined to assist agency staff in prioritizing areas to focus quality improvement efforts. This framework considers issues associated with demand, cost, capital, and benefits to assist in determining which detection and practices are given priority.
Table 2. *QADM Decision Framework*

<table>
<thead>
<tr>
<th>Focus</th>
<th>Component</th>
<th>Guiding Questions</th>
<th>Priority Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demand</td>
<td>Magnitude</td>
<td>What is the degree of error associated with student demographics (e.g., assigned grade, school, district; duplicate records; missing data)?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td>Direction</td>
<td>What are the variations in student characteristics across time (e.g., changed Limited English Proficient (LEP) or disabled category)?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What procedures are used to examine score anomalies?</td>
<td>L-M-H</td>
</tr>
<tr>
<td>Cost</td>
<td>Fiscal</td>
<td>What discretionary funds are available for quality management activities?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>When do conflicting time demands allow for error detection?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How can control integration minimize creating new costs?</td>
<td>L-M-H</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td>What method evaluates yearly results with historic patterns prior to final dissemination?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How are training methods being maximized within the agency across the state?</td>
<td>L-M-H</td>
</tr>
<tr>
<td>Capital</td>
<td>Institutional</td>
<td>How is Quality Management Process (QMP) flowcharting used to improve data integrity?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How are business rules, decision and operational logic, and policy alignment evaluated?</td>
<td>L-M-H</td>
</tr>
<tr>
<td>Benefits</td>
<td>Accuracy/ Precision</td>
<td>Which production sequences are replicated by external agents or a neutral third party?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What steps ensure data from the production sequences are comprehensively reflected in all reports?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td>Spillover</td>
<td>How are behavioral changes attributed to the system’s Theory of Action?</td>
<td>L-M-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How is end-user feedback used as a quality review?</td>
<td>L-M-H</td>
</tr>
</tbody>
</table>
DATA INTEGRITY

Improving Data Integrity using Validation Procedures

Data integrity examines the validity by which inferences can be drawn from the selected data. Error associated with faulty logic, human mistakes, measurement imprecision, and unexplained random error reduces the credibility of targeted outputs. Systematic (faulty algorithms) and nonrandom error (data manipulation) can be particularly troublesome if they occur within federal programs. As a result, there is increasing emphasis on monitoring subgroup, school, and district performance over time. This brings additional challenges to building, maintaining, and utilizing student-level data. All of the obstacles encountered when examining school performance are magnified when considering performance over time. While longitudinal analyses can provide substantively important information, data decision rules made in any one year can have consequences when examining multiple years. For example, student scores that are attributable to school only if the student is enrolled in the school a sufficient number of days might be excluded altogether from multiple year analyses—even though the student has taken the targeted assessments, gains for that student will not exist. There are three potential sources of error when working with longitudinal data.

The first error source is that some student background information is re-entered into the system. In any single year, student information may be internally consistent (i.e., a student is male in every database), but may change from one year to the next (and again be internally consistent). Not only must districts and states have checks to ensure that data are correct over time, but perhaps more difficult is to have rules that direct automated controls to inconsistent entries.

The second difficulty is that while some student background information logically remains constant from year to year, other student information will change over time (although it may be internally consistent in any single year). An example from California is student English language development (ELD) level and language status. Students may advance from ELD level 1 to ELD level 3 from one year to the next and this may appear acceptable; however, it may exist because state or district rules maintain that students can change only one level per year.

The third potential error source associated with longitudinal data is more difficult because in any given year, decision rules can effect inferences about student performance over time. Accounting rules potentially effect inferences when data generated for accountability purposes are used for programmatic evaluation. For example, if students with disabilities (SWD) who receive accommodations have their scores recoded, this may impact inferences regarding SWD academic performance. If SWD scores are recoded from the observed score to Far Below Basic (FBB) to reflect NCLB rules, then changes in SWD performance will be confounded between actual changes in performance and changes in accommodations. Two invalid inferences result from this recoding: (a) more SWD are thought to have improved to Basic and Above (BaA), and (b) the gap between SWD and their non-disabled classmates is decreasing. Table 3 presents results of an examination into the effects of recoding. It should be noted that a previous analyses on these students indicate accommodations on average do not statistically change a SWD performance on the assessments.
Using data from Table 3, the first point to consider is the number of SWD whose scores are recoded from the observed score to FBB in the previous and current years. Considering the overall count of SWD has not appreciably changed, the number of SWD recoded is significantly lower in school year (SY) 2004–2005 than in SY 2003–2004. Among those students whose scores have been recoded, their pre-recoding performance has not changed between the two years. However, the rate of recoding has changed dramatically. Given accommodations (as administered to this sample of students) have little effect; removing them is a step in the right direction. However, inferences that SWD academic performance has improved and that SWD are closing the gap are incorrect. As the forth panel of Table 3 indicates, the 1.9 and 1.3 percentage point change for SWD who are BaA is attributable to changes in testing conditions, rather than changes in performance. In terms of performance increases, this suggests ELA improved only a quarter as much as the uncorrected results imply. For mathematics, the results are more skewed as SWD performed slightly worse than the previous year. In terms of closing the achievement gap, the data lead to invalid inferences because the increase in SWD academic performance is actually attributable to recoding. When considering only that portion attributable to changes in performance, the data suggest the achievement gap between non-disabled and SWD is increasing. Hence, while accountability rules may be based on required federal regulations, these rules need to be carefully taken under consideration when analyses are conducted over time and further when results of these analyses are meant to inform stakeholders concerning school performance.
2.0 Data Quality

The Data Quality section focuses on exploring validation procedures used to detect and control errors within local and state agencies. This section addresses

- documented procedures
- undocumented procedures
- economically disadvantaged
- English-language learners
- students with disabilities
- student demographics, membership, and enrollment
- assessment results

Guiding Question 2.1 What is your state doing to document accountability processes in a written and/or unwritten format?

2.1.1 Documented procedures

Delaware uses a series of standardized written quality control methods to screen data results from its Delaware Student Testing Program (DSTP), Delaware Student Information System (DELSIS), and Delaware School and District Accountability System. Data inputs associated with student demographics and characteristics are cross-referenced with assessment results. This process ensures student participation and performance results have been validated prior to accountability score production and public reporting. The agency is currently augmenting its accountability production process to include multi-year data screening, standardized web-report auditing, and targeted score replication.

Table 4. Sample Quality Control Procedures for the 2004 DSTP

<table>
<thead>
<tr>
<th>Quality Control for the Preliminary DSTP Data Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Receive the preliminary data file from Harcourt with raw scores only</td>
</tr>
<tr>
<td>☐ Initial loading of raw data from Harcourt’s data file</td>
</tr>
<tr>
<td>☐ Check and correct duplicated cases</td>
</tr>
<tr>
<td>☐ Check and correct bad IDs</td>
</tr>
<tr>
<td>☐ Correct wrong information, such as name, gender, race</td>
</tr>
<tr>
<td>☐ Cross check accommodation data with the accommodation database</td>
</tr>
<tr>
<td>☐ Set accommodation flags for those that are not already set</td>
</tr>
<tr>
<td>☐ Set accommodation flags for special accommodated students</td>
</tr>
<tr>
<td>☐ Return the data files to Harcourt</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Control for 2004 DSTP-I and DSTP-II Data Files in General</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Assign aggregation flags according to the aggregation rules</td>
</tr>
<tr>
<td>☐ Merge category data (LEP, Special-Ed, etc.) with the DSTP data from DELSIS snapshots</td>
</tr>
<tr>
<td>☐ Cross checking the raw score and scale score ranges and the consistency of the test scores with the proficiency levels for DSTP-I and the cut-score ranges for DSTP-II</td>
</tr>
<tr>
<td>☐ Prepare the data for the data warehouse and the online report databases</td>
</tr>
<tr>
<td>☐ Convert instructional needs indicator data and load into the data warehouse and the online report databases</td>
</tr>
<tr>
<td>☐ Prepare a SPSS file for assessment</td>
</tr>
<tr>
<td>☐ Calculate state, district, and school level summaries and disaggregated summaries for checking against Harcourt’s numbers</td>
</tr>
</tbody>
</table>

The Hawaii Department of Education produces several types of documentation annually. These include flowchart diagrams used to process data for determining AYP and sanction status, a Functional...
Requirements manual, a Systems Design document, a Reporting System User’s Guide, along with a technical manual for the state’s accountability and reporting system. The accountability manual includes documentation of those internal validity measures used to support AYP determinations. The agency also established the Data Quality Improvement Project in 2004–2005. With this project, a data improvement team was organized to identify priority areas needing improvement. For Hawaii, the data quality procedures associated with English language learners (ELL) and student rostering were the data team’s first areas of focus.

2.1.2 Undocumented procedures

Most agencies have numerous procedures used to ensure the integrity of data used within their information, assessment, and accountability systems; however, many are often undocumented events. Typically, time constraints and limited resources hinder officials from having comprehensive recorded evidence by which to standardized, replicate, audit, and improve their quality assurance practices. In Louisiana, the agency is documenting each step used to validate the accountability data sets and calculations used to generate the indicators. Further, the state is documenting its screening procedures used during score production and AYP determinations. A new set of quality procedures will be established after completion of meetings with the vendor responsible for the testing program. The agency also has a unit of four individuals, whose work is defined in state policy, to investigate unusual data results and review testing accommodations for special needs groups.

In another state, agency staff members have created independent SAS programs to replicate the programming done by the computer department. This procedure ensures all calculations within the programmer’s code matches the business rules as intended by policymakers. Beginning at the individual student level, the SAS programs follow the design logic to produce subgroup, school, and district aggregated data. Further, the replication process functions as an internal control by flagging outlier values using both cross-sectional and trend information. The agency intends to continue refining this control process as the student identifiers become fully implemented.

Guiding Question 2.2 What are some of the unique challenges associated with federal program eligibility/membership?

2.2.1 Economically disadvantaged (Title I)

Those students who are economically disadvantaged are often provided additional resources under those provisions detailed in NCLB, Title I, Part A. This federal program requires state and local officials to maintain extensive documentation of those expenditures paid for with these funds. In Maine, the agency examines the distribution of economically disadvantaged students associated with the Maine Educational Assessment (MEA) results before the information is shipped to the vendor. Maine’s accountability results are produced by the assessment vendor, who has implemented quality control procedures to validate data associated with economically disadvantaged students. Before and during score production, data anomalies are investigated by staff members, who identify suspect data and make any necessary corrections prior to report production. Anomalies are flagged using non-quantitative methods that examine historical membership patterns. Further, isolated island and rural schools with small n-counts are critically reviewed to ensure the business rules are applied correctly.

2.2.2 English-language learners (Title III)

Now funded within Title III, English-language learners (ELL) are becoming increasingly the focus of state and national attention. This phenomenon is partly a result of the large number of non-English speakers entering school systems in states. Mexican or Latin Americans, the fastest growing minority population within the country, are bringing attention to those federal dollars allocated under Title III. The diversity within the ELL subpopulation has required states to implement complex coding
systems for inclusion into accountability and to monitor the acquisition of the English language.

**Delaware** uses a series of nine acronyms and associated business rules in making program and accountability determinations. The Delaware Student Information System (DELSIS) uses a coding structure to separate the different ELL categories. These data are cross-validated by program staff within the agency to ensure local school districts apply the correct definition to those data being input into DELSIS. **Oregon** produces preliminary AYP reports and other draft reports prior to the release date. The agency flags cells as “no data” for any school or school district that does not submit the required AYP information by the due date needed for report production. Districts and schools with “no data” for any rating are required to implement corrective actions; however, if required data submissions are not received for final AYP determinations, the business rules require all “no data” designations to be reclassified as not meeting AYP. These rules are to prevent the exclusion of students with limited English proficiency from AYP determinations.

### 2.2.3 Students with disabilities (IDEA)

The SWD subgroup continue to be a data quality area challenging state and local agencies. Although the IDEA provides important legislative rights for parents and students with disabilities, each state has its own programmatic guidelines for inclusion, access to the general curriculum, assessment participation, testing accommodations, and identification procedures. Operational definition must reflect what tests and how students participate in a state’s assessment system. Further, these definitions must be clearly documented within both policy and testing guidelines promulgated by the state and/or district.

In **Delaware and Louisiana**, the agency examines the change in the SWD performance data for those students participating in the statewide assessment program. Multi-level analyses are conducted to detect distribution pattern changes across the state as compared with previous trend data. Qualitative, onsite monitoring procedures coordinated across the agency guide validation inquiries into the atypical distributions. These efforts are conducted as an augmentation of the state’s consolidated monitoring of its federal programs. **Delaware** has implemented a quantitative examination of the distribution patterns across different entities with a focus on changes in accommodations assigned across different grade levels. This control procedure, detailed in the state’s accountability manual (DDOE, 2005) will add an automated detection component for both accountability and assessment systems.

One challenge in determining the graduation rates for SWD is that in some cases the students exit special education prior to the end-of-year collection date. These membership data change the denominator needed to calculate the four year cohort rate used in **West Virginia**. The state is currently examining ways to control for these mid-year changes in a manner by which the aggregation rules adjust for the data changes. In a similar dilemma, **Delaware** and **Pennsylvania** experienced high school students being promoted across the selected assessment grade. This phenomenon can exist when local pupil progression plans allow for students to remain in one grade as a repeater, continue to earn credit hours, and then be promoted at the mid-semester. As a corrective action in Delaware, the agency requires all students participate in the 10th grade assessment regardless of whether they advanced past the grade. **Kentucky** addressed this assessment issues by requiring students to participate in the levels-based assessment, with the only exception being 12th grade students planning to graduate early. In these cases, the student submits a portfolio rather than completing the on-demand writing assessment.

---

**Guiding Question 2.3** *What validation procedures are being used to improve data quality and address unique challenges in your state?*

### 2.3.1 Student demographics, membership, and enrollment

State and local education agencies typically implement data field specification within their information management system as one error detection and control procedure. This process rejects data
input by end-users that does not match the pre defined field requirements. Also, most information systems have auditing programs running either currently or at predetermined intervals to flag problematic data. End-users receive syntax or error messages notifying them of the data integrity issue, thus prompting corrective actions before the record is accepted.

**Delaware**’s *SingleSignOn*, web-based portal allows its school districts to examine and validate aggregate demographic data used for AYP determinations. Using preliminary data, district and school building administrators are then able to drill down to the student level to view student demographic data comprising the subgroup. This process allows for data errors to be corrected prior to the final accountability results being produced. As a result, score appeals due to data anomalies have significantly decreased since the agency implemented its web-based, validation system.

**Nebraska**’s web-based information portal is used in conjunction with its District Assessment Portfolio process. Here districts enter specific information regarding those building and subgroups within their jurisdiction. The agency monitors and provides feedback regarding those membership data being entered into the system. Selected business rules are provided within the web pages to assist end-users during the data entry and examination process. The agency uses accountability membership information received during the online data collection process to guide onsite monitoring.

For **West Virginia**, operating a centralized student information system, students assigned to alternative schools or programs on other campuses are routed back to their school of residence. This policy assists in preventing the loss of data, while encouraging resident schools to monitor the educational services being provided in the alternate setting. For alternative students moving from one district’s program to another, the assignment of resident schools has become more problematic, often resulting in data being aggregated to the next higher unit of analyses (i.e., district and state). Conversely, some states are addressing students having multiple resident schools and districts. In **Oregon**, duplicate state student identifiers are flagged by the agency. These records, located within the *Secure Assessment* application are matched using specified criteria (e.g., last name, birth date, gender) then presents to the end-user on the screen of the resident district(s) requesting the identification number. If the districts verify that two numbers have been requested for a single student, the state merges the student under the lower of the two numbers and closes the duplicate number through an automated process. This process controls the enrollment error of assigning accountability data to two or more educational entities.

### 2.3.2 Assessment results

Assessment data are a major input into the production of accountability results. For statewide assessment systems, most contractors implement quality assurance measures within their contractual agreement. Many of these processes are generically described in assessment technical manuals because of proprietary rights to the procedures.

In **Maine**, staff members travel to the vendor’s location and conduct a series of quality checks to ensure the data being used for assessment and accountability results are accurate. For example, the vendor conducts parallel processing during score production by which separate computer programs are written to replicate the final results. Hand-scoring a representative sample to validate each report cell is conducted as outlined in the Appendix E, MEA Technical Manual 2003–2004 (MDE, 2004). Similarly, in **Delaware**, the assessment, accountability, and information management staff work closely to examine those data being sent to the vendor prior to score production. Then, key staff members implement a standardized quality control checklist (see Appendix A) prior to publication of individual, school, and district results. The results of these efforts have reduced the risk of nonrandom error entering the production process; however, these procedures are limited in scope: the quality control efforts are limited to information management and assessment inputs, but are not extended into accountability score production. This exists because the agency implements its accountability policy using internal resources, including programmers and third-party contractors. Accountability error detection and control procedures follow a standardized, documented process like those used in the
assessments system. Both Maine and Delaware are able to provide evidence to federal reviewers that ensure results from the system are being validated prior to public release, including data from their alternate assessment programs.

**Louisiana** conducts parallel replication with its vendor during the score production process. One error detection procedure conducted for the agency by the vendor is to flag test records with unusual wrong-to-right distribution (erasure analysis). These data, along with writing samples having highly similar structure, are provided to local officials for further action; however, the agency does conduct onsite investigations in extreme cases. Further, the agency sends trained observers to ensure the standardized administration of the assessment system is actualized within classrooms. In the future, the agency will be able to identify a testing subgroup with remarkably similar test responses. The vendor’s addition of a testing group identifier to the answer documents will assist in auditing activities by the agency.

| Guiding Question 2.4 | What experiences do you have to share that would help others? |

### 2.4.1 Lessons learned

Communication and collaboration are critical for a multi-faceted program such as NCLB. In 2003, in **Hawaii**, the state education agency (SEA) instituted two critical changes impacting the identification of students with special needs. A new data collection instrument was developed specifically to collect and manage records for the LEP student population. Initially, this proved to be extremely difficult in both ensuring data quality reviews and coordinating the timely submission of data for school accountability purposes. In special education, a modification was made to the codes used to identify students as special education or pending certification. Unfortunately, the codification changes did not reach the program staff responsible for school accountability and student assessment, causing data irregularities within several subpopulations. In the aftermath of this communication delay, the agency now conducts NCLB AYP Debriefing meetings for key staff members who are responsible and contribute to AYP score production. These meetings have helped to highlight the past year’s accomplishments and challenges while identifying quality improvement efforts for the upcoming year.

### 2.4.2 Caveats

**Maine, Nebraska, Oklahoma, and Utah** have emerging or newly established student information systems. Student level information associated with federal program membership from previous collection structures is being migrated into a comprehensive system. Screening data inputs using these newly established systems will continue to identify data errors and inconsistent, multi-year distributions. One caveat for emerging information management systems is to break the cycle of inaccurate data by documenting error detection procedures and implementing training frameworks to build end-users capacity. In regard to auditing, **Louisiana** suggests not wasting time asking test administrators if they knowingly compromised the data’s integrity—no matter how the question is asked, they know what they should say: No. Put policy in place that requires the politicians to agree to stop some punitive action (e.g., voiding scores) rather than have them decide whether or not to initiate corrective actions.
END-USER FEEDBACK
Using Feedback to Guide Training and Determine Impact

Each state’s unique context and educational initiatives shape the method used to build capacity both within the agency and throughout local educational agencies. Nonrandom error enters accountability systems when end-users, who provide critical information, misinterpret or misrepresent actual events within schools. Vague or poorly written business rules for data collection produce results that don’t represent the students or schools. Typically, technical training and onsite support can reduce the presence of these errors prior to the production of accountability results. Limitations in sustaining a skill level across an entire state in an evolving accountability design are both time and labor intensive activities. Many states have focused on increasing the capacity of key central office staff to ensure that all end-users have a comprehensive understanding of the numerous business rules and evolving policy. Although a daunting task, as LEA staff use their own data feedback structures to ensure accurate accountability data, nonrandom errors will significantly reduce.

Accountability system results are designed to produce intended consequences such as recognizing schools and districts demonstrating best practices in the areas of instruction, curriculum, and assessment. Validation methods provide a process used to determine the quality characteristics, which can be modeled and replicated. The causal link between intended consequences and behavioral is fraught with assumptions, which are difficult to evaluate.

Another problem with placing too much emphasis on the school level is the ecological fallacy (i.e., making inferences about individuals based on group data). For example, a school A having lower average performance than another school B, but having a higher proportion of ELL students (60 percent for school A vs. 30 percent for school B) implies that the more ELL students a school has, the lower mean performance will be because ELL students do not score as high as non-ELL students. By using school averages we do not know whether low average scores are truly the result of lower performance by ELL students or other factors (e.g., very low performance by non-ELL students in school A).

3.0 End-User Feedback
The End-User Feedback section focuses on understanding how information from end-users can guide quality control development, refinement, and improvement. This section addresses

- training format and focus
- targeted audiences
- model practices
- staff mobility and sustainability
- customized training
- modular materials
- policy sanctions

Guiding Question 3.1 What is your state doing to build capacity at the local level to minimize data errors and accountability irregularities?

3.1.1 Training format and focus
Numerous training formats are used to build capacity within a state, including regional trainings, train-the-trainers, video, and large-scale information sessions. Although some small group and customized trainings are implemented for the expressed purpose of improving error detection and control procedures within information management, assessment, and accountability systems. A heavy
emphasis has been placed on the administration of statewide assessment systems and the screening procedures used to validate score results. For example, Oregon posts a schedule of due dates for data collections prior to the start of each school year. Business rules, validation procedures, data definitions, and other accountability information are available on the agency’s website. Further, this information is available for prior years with updates for the current year posted at least a month in advance. The agency conducts training associated with these due dates for accountability information via teleconferences. These videoconferences are delivered live using interactive video, and then archived for later viewing by central office staff. In addition, training of SEA staff was addressed by developing a data owners group. This group provided training along with those primary responsibilities found within their job description. The state continues to focus on building capacity at all levels necessary to support its assessment and accountability systems.

Nebraska has placed building the assessment and accountability literacy capacity of teachers and administrators as the keystone to improving student achievement for all children. Professional development in the areas of standards alignment, assessment literacy, assessment quality, data analysis, and the school improvement process continues to be a priority for teachers and administrators. Other support reaches Nebraska’s schools in a variety of ways. Technical materials, hands-on workshops, conferences, statewide information sharing sessions, interactive databases, onsite technical assistance, and training sessions conducted by educational services units buttress the improvement efforts of local assessment and instructional practices. The agency builds local capacity through and in developing, administering, and aligning local assessments, often through consortia of several districts working together. The agency and its service center partners focus on identifying and reducing assessment and accountability error at the data source, thus reducing the need for ex post facto data audits. Onsite visitations by agency staff have provided critical end-user feedback about the assessment quality, student membership, and score irregularities.

3.1.2 Targeted audiences

Each state has key audiences to which it directs a majority of its training and support efforts. District and school personnel, especially in small, rural school districts, are frequently required to attend numerous workshops and trainings conducted by state agents. Often these local personnel are principals and key administrators who must balance daily demands with those of external parties, such as parents, school boards, and state officials. These educators and information managers are rarely the actual individuals entering and reporting student-level demographic, enrollment, or membership data. The actual end-users, often secretarial or clerical staff, must interact with a perceived never ending set of complex rules. Unfortunately, this audience rarely participates in state-level trainings that focus on managing and quality controlling data they themselves input into the system. Rather, the burden for quality assurance falls on school and district administrators.

In one district, dropout data sent to the state continued to be plagued with end-users’ input errors. Although the state produced validation reports six times prior to sending the data into the accountability score production process, the impacted principals were unaware of the reports. As a result, the accountability scores contained nonrandom dropout data because the end-users who attended the mandatory state training failed to manage the delegated task of data correction. To ensure these results were reflective of actual events within the school, the district implemented random data audits. These data were discussed with the principals, along with suspect coding errors and projected end-of-year performance. The error detection process used quantitative, non-documented, procedures to focus the staff training in understanding sources of error, then management techniques to reduce their presence. In the following year, the district was flagged by the state for unusual improvement in their dropout statistics and was audited by state officials. Had the district documented its detection and control procedures, the state officials would have been able to conduct a desk audit rather than spend three days in the district.
Guiding Question 3.2  What are some of the issues around capacity building at the district/school level that are producing unique accountability challenges?

3.2.1  Staff mobility

All schools and school districts have some degree of staff mobility, either through attrition caused by retirement, exiting the profession, or simply transferring to another educational entity. When mission critical personnel leave, so does the human capital they possess, including the job experience and numerous training opportunities. This phenomenon is clearly present within the information management and analytical/psychometric field. End-user training focused on error detection and control requires individuals with technical skills that are often developed through state efforts are in high demand. For example, in one large metropolitan district, the data manager quit during a major system conversion. This forced the emphasis to getting the results out before examining whether the data were accurate because no replacement could be found with the necessary skill set. Maine has numerous small, rural districts in which only one or two key persons have a comprehensive understanding of the Maine Learning Results data used for the accountability system. The department provides training materials; however, when key central office members exit the system, seldom do replacements have an adequate understanding of the accountability design. This limitation is prevalent in other states with numerous small districts, such as those in Montana, Nebraska, Pennsylvania, and West Virginia.

3.2.2  Sustainability

The training demands associated with sustainability are significant in light of the continuous changes to the business rules used within accountability systems. Compounded by changes in other parts of the system such as new assessments, data reporting requirements, and program monitoring, the ability to maintain end-users with up-to-date information at a skill level necessary to examine current practices and formulate improvement efforts is significant. Unfortunately, end-users must often focus on the basic operations of the task and rarely have the time to reflect on new practices in an effort to improve efficiency. Training materials are in the constant work-in-progress mode. Further, efforts to build foundational skills for new staff and expand those with experience require a differentiated instructional model, which are difficult for large scale efforts.

In Illinois, the agency produces information about its accountability system and promulgates information via its website, along with most state agencies. The continuous reduction of staff, along with a lack of regional support structures, has reduced the training opportunities necessary to fully examine detail accountability data. Louisiana conducts yearly updates on how to accurately label test documents and properly submit attendance and dropout data. Local officials are given multiple opportunities to review and correct data prior to its use in accountability results. The agency sends comprehensive data sets used to calculate the accountability scores to local officials prior to release of accountability results. The agency also provides spreadsheet-based calculators and extensive training necessary for district staff members to replicate accountability calculations and determine impact of data corrections. However, the increasing complexity of the accountability system has limited the number of individuals at a local level who comprehensively understand the system.
Guiding Question 3.3 What support techniques are being used to reduce data irregularities? Are they replicable for other states?

3.3.1 Customized training

Assessment training has been customized for school and district administrators in an effort to build local leadership’s ability to detect data errors associated with the implementation of Nebraska’s assessment and accountability system. The commissioner of education worked with representatives from 17 institutions of higher education to develop an 18 month series of assessment leadership workshops. The Nebraska Assessment Cohort placed emphasis on improving understanding of classroom assessment practices, while the Leadership for Learning program for school leaders or curriculum directors stressed improving local assessment systems. After the initial Leadership for Learning cohort finished its five-week session in SY 2002–03, a third party evaluated the course to examine ways to improve the training. Additional teacher education assessment projects strengthen the partnerships with higher education.

3.3.2 Modularized materials

Louisiana developed a series of standardized training materials to build local capacity. The focus of these modules ranged from understanding the design logic and business rules to developing school improvement plans. Policy required members of the each school’s improvement team to participate in the initial five-day training, then subsequent annual update trainings. A localized database maintained documentation of the members who had completed the training requirements. These standardized modules provide continuity across the agency and region service center teams. Changes to the accountability system, along with training on how to examine data being used to produce accountability results continued through the NCLB changes.

3.3.3 Policy sanctions

In Ohio, state statute requires that all students participate in the regular assessment, assessment with accommodations, or an alternate assessment. Districts failing to report all enrolled students in a tested grade to the information management system are subject to fiscal penalties. One penalty is for the withholding of state foundation payments for those students missing test records. Conversely, in another state, the board of education has been unsupportive in efforts to require better response from the local school districts. Policy sanctions are often not enforced, although the state board did recently approve policies necessary to transition the assessment system and graduation calculations into compliance with NCLB. These consequences are meant for the expressed purpose of increasing the accuracy and precision of accountability results.

Guiding Question 3.4 What experiences do you have to share that would help others?

3.4.1 Lessons learned

Nebraska has placed significant emphasis on improving the communication and partnership to and from stakeholders. Unlike passive reporting systems, the state has developed a multi-layered, interactive communication structure by which educators and the general public receive and provide information about the assessment and accountability systems. Distance learning opportunities using satellite broadcasts, such as Chats with Pat and Dot, provide state educators with real-time question and answer sessions about data and business rules used within the systems. These opportunities assist in promulgating long-term improvement plans, changes to the accountability design, and associated business rules. Here, the agency uses feedback from local schools and school districts to determine the degree by which intended consequences are being actualized across the state.
3.4.2 Caveats

Louisiana developed district assistance teams (DAT) using modular training materials. DAT members were to assist low performing schools in examining assessment and accountability data to develop improvement plans. DAT training activities reduced the number of appeals while aligning improvement activities to the accountability system. By 2001, the agency received feedback about limited personnel available to support the growing number of low performing schools. The districts claimed they neither had the personnel nor time to complete job tasks while continuously supporting school improvement activities. The proactive approach to improving accountability results continues, but accountability design changes have resulted in few local officials fully understanding the system.
CREDIBILITY

Promoting Quality through Mapping and Replicating Production

Credibility of an accountability system is improved when systematic and nonrandom errors are identified and controlled. Producing accountability scores requires converting federal, state, and local policies into a coherent set of business rules within the system design schematics. Business rules provide the operational details necessary to convert these broad policies into a clearly defined set of steps. These steps organize data elements into a logical sequence necessary to produce selected results, while screening for unwanted error. Translating policy into business rules that can be unambiguously implemented is a challenge because it often requires people with different perspectives on accountability (e.g., legislators, administrators, state-level program specialists, and computer programmers) to understand what is needed to address the design logic.

Several quality control procedures serve to ensure production results are credible in that they have been comprehensively examined for nonrandom error. These control mechanisms can be deployed to follow detailed production mapping and/or simply use professional judgment to examine the resultants for aberrant data.

Regardless of the approach an agency selects, the overall strategy is to establish a reasonable, cost-effective set of controls that ensure the accountability results are credible. This should not be misinterpreted as set of minimum criteria, which once attained require no further action, but rather as a dynamic process of quality improvement. This process should build on prior experiences, establish action plans for future improvement, and strive to reduce costs (fiscal, human, time) through process streamlining and automation. These actions promote organizational efficiency while strengthening the quality of accountability results.

4.0 Credibility

The credibility section focuses on the quality assurance procedures used during and after the production process, with emphasis on mapping the system’s design logic and business rules in conjunction with production controls. This section addresses

- technical manuals
- standardized controls
- professional screening and examinations
- programming and non-standard reviews
- time constraints
- adequate staffing

For Consideration...

District staff members are frustrated because so much of their time and resources must be spent on providing the department with data. Department staff who are responsible for programs have commented about the increased amount of work and that the work often requires skills they do not have, or are not interested in acquiring. Programmers are also frustrated because they have too many tasks without the sufficient prioritization from management to complete essential tasks. A major newspaper filed a freedom of information act request because they could not understand why a report was coming out later than it had in the past.

How can so many people agree that good information is essential and also be so upset about how the data are being collected, validated, analyzed, and reported?
Guiding Question 4.1  *What is your state doing to examine and validate accountability results during the score production process?*

4.1.1  **Technical manuals**

Accountability technical manuals have not become a standard practice in most state agencies and are very rarely published with local accountability systems, although Dade County, Florida, and St. Tammany Parish, Louisiana, are two notable exceptions. For assessment systems, technical manuals provide the reader with information about the assessment’s design, technical properties, administration, scoring, and reporting formats. For accountability systems, a comprehensive set of procedures, business rules, and formulas must also be documented so they are readily available to the general public, compliance reviewers, and educators. This information about the technical procedures used to assign accountability judgments on local schools and districts is critical in promoting confidence with the system, while concurrently developing quality assurance procedures that promote efficient and effective resource allocations.

**Ohio** provides stakeholders two key reference guides, one for building-level accountability determinations and the other for districts. These documents provide detail information to educators and the general public about the methods used in making AYP determinations and publishing local report cards. The agency organizes the technical information into standardized modules detailing the components of each indicator. For example, for high school graduation rates, the business rules are presented with clarifying narratives, computational formulas are reported along with data elements and sources. Further, the design logic is communicated within the filtering section, thus helping the reader understand conditional relationships. Likewise, **Nebraska** has provided the decision logic in both narrative and flowchart formats. Both states are readily able to isolate components within their accountability system in which unwanted error may occur.

Other states, such as **Arizona** and **Oregon**, have developed and organized their technical manuals around key accountability components. Common features within both documents are information about the data types, business rules for key accountability indicators, and how to determine overall subgroup, school, district, and state AYP status. The **Texas** Education Agency has published a technical manual organized around three major components: (a) standard procedures, (b) alternate education accountability (AEA) procedures, and (c) items common to standard and AEA procedures. Within each section, the agency documents the performance standard along with the method and data used to determine if the entity of focus has met the annual objective. Other information, business rules, and report interpretation examples are used to assist the reader.

4.1.2  **Standardized controls**

The implementation of standardized controls allows for evaluations to determine if the action resulted in the desired effect, the reduction in the magnitude and direction of nonrandom error. These defined procedures are useful within and outside of an agency, in that they allow third-party replication and/or auditing function by which the conclusions reached by the accountability design can be validated. The earlier in the production process that input data are placed through a rigorous, standardized control process and deemed valid, the lower the probability of invalid accountability consequences being applied to schools and districts. Many data integrity issues can be automated into data collection processes. For example, the student records in files submitted to preprint answer sheets can be checked for correct field specifications (i.e., required information included, dates of birth in range, etc.), then cross-referenced with other information to validate the entry such as rejecting duplicate entries. In **Louisiana**, the agency is standardizing the process used to check accountability data and the decision logic.
Score production typically moves into a report production process that aggregates and filters data according to prescribed business rules. In Delaware, the agency converted its non-standardized, professional screening procedures associated with its on line, accountability reporting system to a documented procedure. A third-party contractor audits accountability results using a stratified random sampling framework. A conscious effort is made to include large and small schools, new schools, and idiosyncratic situations such as schools in a district that administer Department of Defense schools. Generally, 16 to 20 schools are audited, which is approximately 10 percent of the population. However, if problems are encountered, additional schools are checked and rechecked until the quality control manager is satisfied that the problem has been corrected. The control procedure is basically a process of (a) checking current year computations and the internal consistency within the school’s screen set, (b) checking annual progress including confirming the previous year’s data, and (c) checking school improvement for the past three years. Table 5 provides an excerpt from the standardized procedure (see Appendix B for the entire sequence).

Table 5. *Accountability Score Quality Control Procedures*

<table>
<thead>
<tr>
<th>Participation Target Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Review number of cells that are eligible for participation review. Cells with &lt;40 students are excluded. [screen 2]. Six cells for each subject are eligible.</td>
</tr>
<tr>
<td>b. Check calculations for appropriate cells, cells with large font. 95 percent participation of eligible students is required.</td>
</tr>
<tr>
<td>c. Confirm that data agree with the status as listed at the top of the screen, meets or does not meet participation.</td>
</tr>
<tr>
<td>d. Confirm that [screen 3] information agrees with this screen, [screen 2].</td>
</tr>
</tbody>
</table>

### 4.1.3 Professional screening and examinations

Staff members with program knowledge and experience can be used effectively to screen the production processes and evaluate accountability reports to detect errors. However, some data checks require more expertise than others. For example, it may be useful to have program specialists (e.g., for ELL or SWD) check that sufficient numbers of students in specific subgroups are being administered tests. In cases where the exact figures to be submitted are not known, comparing the data to statistical estimates of what is expected can identify problems. For example, each school’s enrollment roster can be compared to the previous year’s enrollment so unusual changes can be detected. In this case, differences of more than a certain percentage could indicate a data problem and prompt control activities.

Another approach used in Nebraska is to flowchart the design logic used in making judgments. This flowcharting process allows less technical and policy oriented staff to evaluate if the business rules are being applied as intended by the policymakers. Figure 4, from Nebraska’s Assessment and Accountability Manual (NDE, 2004) provides an example of how professional examination of the decision logic does not require examining the programmer’s code. Through the application of flowcharting to assist in examining the review process, the state is able to focus on potential error entry points and implement the appropriate controls.
4.1.4 Programming reviews

It is important to take the time to ensure that program code implements the policy decisions. Translating policy into business rules is an iterative process. Do not assume that once the business rules have been written and the coding completed that the process is complete. In Delaware, each module is examined and documented within the technical manual. Further, the code is reviewed to ensure that policy intentions are being implemented and data are run through the system to ensure the program is complying with the business rules. The business rule author and a third-party examine the results to validate the code is operating correctly. Nebraska contracts the accountability production process to an outside vendor. As one quality control, each production module used by the programmers is flowcharted with supporting narrative explaining the operational logic being used to execute the necessary code. These diagrams are then professionally screened by the agency to ensure they reflect the design logic and policy intent. Figure 5 was taken from the agency’s technical manual (NDE, 2005).
4.1.5 Non-standardized reviews

Delaware and Oregon provide districts with preliminary AYP reports and report card assessment data counts prior to public release. These data assist districts in identifying data quality issues associated with AYP determinations along with verifying the subgroup, school, and district accountability scores. Louisiana takes a sample of schools and districts to audit based on preliminary results. Using the original data, the sampled entities track it through the system to make sure that the ratings are being correctly calculated. Additionally, Maine and Oregon provide spreadsheets in which schools and districts can enter data to predict official determination and further examine score results. Maine samples a large municipal district, a school administrative district, a union, and a very small district. The process uses the actual data printed in the entity’s assessment report to examine the AYP report as a quality assurance step ensuring correct implementation of AYP decision rules. The agency then compares the results with the vendor to check for any inconsistencies or potential reporting issues. These procedures are documented to increase standardization and for inclusion in the state’s accountability manual, currently being developed by the agency.
Guiding Question 4.2  What are some of the issues associated with checking score production during the process?

4.2.1  Time constraints
For many, if not most, states, tests are administered under tight timelines. Therefore it may be difficult to build in time for extra checks on data integrity. In one state, the timelines outlined in NCLB along with media requests for the agency to publish reports on pre-defined dates along with additional and clarifying information time has increased pressure for high quality data. Backward planning has assisted the agency in attaining established reporting objectives; however, many quality assurance efforts are compromised for the sake of expediency and self-preservation. These time constraints are often exacerbated by districts claiming the data contain errors and that the assigned judgments are unwarranted. Some states, like Texas, do not allow schools and districts to appeal accountability judgments on the ground of data error. The rationale for this policy is based on the principle that the district is responsible for the data by which accountability scores are produced. Time constraints are perhaps the most divisive issues surrounding error detection and control at the district and state levels.

4.2.2  Adequate staffing
The human resource component is critical when establishing error detection and control activities. Inadequate staff to complete the essential production and program task is typically a concurrent limitation with ongoing time constraints. In one state, the staff tasked to conduct data integrity checks was reallocated to produce test materials and assist in test administration. In Louisiana, accountability policy evolves constantly and requires almost weekly discussions on what needs to change in the application of the policy to calculations and decisions. The agency is very efficient at making changes, but rarely gets a respite from such tasks to engage in capacity building activities across the state.

Guiding Question 4.3  What types of replication checks have been successful in reducing unwanted errors? Are they replicable for other states?

4.3.1  Comprehensive vs. Targeted
The design logic associated with making NCLB-defined accountability decisions has inherent pressure points. Data errors at these points within the system have significant directional influence on AYP determinations. When available resources are limited, a state may choose to target checks for specific errors rather than checking for everything. For example, spending more time and effort to check on the error that caused the most problems the previous year might be more efficient than checking all fields equally. Data validation might focus on districts or data providers that produced the most error the previous year. Instead of analyzing collected data to see if it is accurate, another option is to completely duplicate the process and check for identical results. This can be an expensive option, but for critical data, it provides a high level of confidence and credibility.

In Louisiana, the agency has policy staff having the skills needed to writing programming code that duplicates accountability results. Also, the agency has key programmers who maintain an ongoing understanding of policy changes approved by the state board of education. Louisiana’s ability to execute parallel computations has allowed staff members to proactively identify error sources attempting to infiltrate summary results prior to public release. In Iowa, the AYP algorithms are programmed to run automatically so the agency can use a screening technique to identify nonrandom errors. The agency screens for aberrations in the n-counts and determines through professional judgment if the data are atypical. When anomalies are identified, staff members contact districts to investigate the issue.
Further, the agency duplicates all automated computations associated with Safe Harbor. This procedure validates whether or not the entity attained the targeted AYP thresholds.

4.3.2 Third-party

Some states find it useful to have a contractor that is not responsible for collecting or producing data conduct checks on data accuracy. States with separate contracts for different parts of the test development, administration, scoring, or reporting may arrange for one contractor to duplicate the work of another to provide a data accuracy check.

In Hawaii, the agency conducts replication checks annually using a triangulation method. This triangulation process incorporates work done by the agency, a contractor, and an independent third party. The replication is targeted at AYP processing beyond adjustments to update or correct individual student-level records. In other words, the agency replicates score production after input data corrections have resolved duplicate, missing, or incomplete student records.

4.3.3 Stakeholder reviews

A good practice is to create an Excel spreadsheet into which district staff can enter their school’s results and have the accountability ratings calculated. This allows many people to try out the calculations with data they know and care about. Ohio, using their accountability manuals as training tools, has focused on creating the ability of local districts and schools to validate accountability results. These reviews often expose unforeseen problems within rating calculations or with unusual and unexpected situations by which the business rules may not address. In Louisiana, the agency conducts monthly workgroup meeting with the various staff involved in accountability processes in the agency. Consultant and upper-level administrators are present to address emerging issues. The agency designs simulations to evaluate the impact of the various implementation choices and reach consensus on how best to proceed.

Guiding Question 4.4 - What experiences/war stories do you have to share (good or bad) that would help others?

4.4.1 Lessons learned

During the preparation for the U.S. Department of Education’s Standards and Assessment Peer Review, one state discovered a programming error that resulted in inaccurate disaggregated proficiency data. These data had been placed into the public arena after having provided each district with a substantial timeframe to examine and verify their AYP information. The design logic appeared valid; however, because students who received accommodations have multiple membership characteristics (students with disabilities who are also economically disadvantaged), the programmed aggregation rules simply followed the algorithm. In this example, the business rules restricting reporting accommodations to the special education were not established. AYP determinations were not impacted because all the students with disabilities were accounted for in the computations. The agency has now documented its business rules and design logic within its technical manual along with implementing random screening procedures.

Three-way replication (triangulation) of AYP and sanction status determination has ensured the algorithms and associated software applications are producing results paralleling Hawaii’s design logic. The strength of triangulation as an internal validity check for AYP determinations was demonstrated early in Hawaii’s NCLB tenure. This approach was successful in identifying a software problem during the 2003 score production period, which was the first year the state procured a contractor to assistance with AYP processing. (Kentucky experienced a similar software conflict between the programming code used within its accountability and assessment system.) The lesson learned by this experience resulted in the agency adopting the triangulation methodology as a standard operating procedure during
score production. The requirement for three-way convergence of independent computations, albeit more difficult and time-consuming, has served well as a quality assurance procedure for accountability determinations.

In Alabama the accountability results and associated judgment were assigned; however, the attendance data for most school districts were fraught with end-user error. The agency worked with local officials but not until the official AYP release did the data clean up become an important consideration by many schools. The resulting deluge of appeals required the agency to recalculate the results with updated attendance results. The subsequent year, the agency relied on its vendor to produce the accountability calculations while the staff members screened data prior to public release. This early detection was also implemented in Louisiana after the release of its 2000 accountability results. A software conversion error had produced attendance values unreflective of actual events on the ground. State officials recalibrated the districts and associated school results and used the updated values in determining the subsequent growth targets.

4.4.2 Caveats

Within the Ohio Department of Education, Office of Policy and Accountability, staff members lead a Local Report Card Data Team. These data teams are comprised of program and information technology staff members across the agency, which allows all business rule changes to be discussed from the programmatic perspective. During these discussions, the information technology staff provides input and reacts to programming demands and data availability. Throughout the year, unique and/or unresolved issues are logged to document any policy changes and the resolution reached by the data team for implementation. The team has an intranet site for communication so members can access to add, edit, or review any issue or decision made in the past year. This interagency problem and resolution strategy was implemented in another agency. Because the unique and complex accountability design required continuous improvement to address the plethora of issues arising from implementation, the agency formed an accountability planning committee. Members represented information systems, accountability, and assessment along with several senior administrators. This group examined issues and formulated solutions, which were examined by a third-party consulting team responsible creating the original design logic. Empirical testing was often conducted to evaluate the impact of policy recommendations, for example, determining the number of schools impacted by rerouting alternative students to their home school. The challenge has been in sustaining the skill level and experience resulting from staff mobility.

In Delaware, the department has taken significant steps in standardizing the quality controls being used throughout the production of accountability results. During the initial development and implement, several external audits (e.g., GAO, Governor’s Executive Order 54) were conducted to examine the validity of the assessment and accountability results. These external queries cost the department considerable time and personnel to produce validity evidence. One document, the Accountability Technical and Operational Manual (DDEOE, 2004; DDEOE, 2005), made available the actual programming code, decision logic, and business rules associated with Delaware’s Accountability System. Since its publication, state officials have identified potential entry points for error and have begun developing both quantitative and procedural mechanisms within the agency to further validate accountability results.
SUMMARY
Looking Back-Looking Forward

This document begins exploring the quality assurance practices used to detect and reduce nonrandom error across three key educational systems: information management, assessment, and accountability. Accountability, along with other systems, must ensure judgments about subgroups, schools, and districts produce credible inferences that guide and motivate action. Using four foundational principles inherent to outstanding public sector organizations, state and local officials are further able to examine their current assurance practices across a developmental continuum. The Quality Assurance Diagnostic Matrix is the first attempt to blueprint those practices being used to control validity threats. Information recorded with the matrix focuses on areas in which error detection and control procedures would improve accountability results, while promoting more efficient operations. As educational entities seek additional ways to prioritize quality assurance development and implementation, descriptors and examples from member states help contextualize the issues.

Numerous quality-enhancing activities are being used to ensure accountability judgments are fair, honest, and open. Capacity building frameworks readily help end-users understand how the numerous business rules operate within the system’s intended design. Unfortunately, many practices remain undocumented and non-standardized, which creates a difficult environment to initiate comprehensive improvement activities. Balkanization, time demands, legislated mandates, and human capitol all contribute in reducing the effort to establish comprehensive quality assurance practices that are ongoing and verifiable. This document details some error detection and control practices along with lessons learned during the implementation of NCLB. Looking into the future, improvement efforts by state and local agencies to streamline activities through planning and monitoring, along with the evaluation of those efforts, will assist in aligning systems. This alignment initiative will focus on promoting a coherent system of continuous improvement, both in production quality and efficiency.
REFERENCES


Nebraska Department of Education (2004). *School-based, Teacher-led Assessment and Reporting System (STARS) validation design.* Lincoln, NE. Author.


Palaich, R. M., Good, D. G., & Van der Ploeg, A. (June 2004). *State education data systems that increase learning and accountability.* Naperville, IL: NCREL.


APPENDIX A
DSTP Quality Control Checklist

1. Download the data file correctly
   - Check the total number of cases by grade
   - Change some variables’ names (no more than eight characteristics and consistent to previous years)

2. Clarify the cut-off scores
   - The aggregation rules due to the SAT9 accommodations
   - The new aggregation rules at the state level
   - The cut-score ranges for DSTP-II

3. Review the 2004 DSTP data documents
   - Review the answer key files
   - Compare the keys with the data layout to locate the core items and field test items
   - Review the score codes for SA and ER items and writing assessment
   - Review the raw-scale scores Conversion Tables to identify the theoretical ranges of scale scores compared with the real score ranges
   - Review the raw-PR-scale tables using the SAT9 norm book by grade

4. Check the accuracy of the 2004 data by grade
   - Run frequency distributions by grade and test (all test scores, aggregates, and accommodations)
   - Check score ranges (record the results in the tables)
   - Check invalid scores’ codes and meanings
   - Check the consistency of invalid scores with score ranges
   - Check the raw-scale score conversions for SBS using the conversion tables
   - Check the raw-PR-scale score conversions for SAT9 using the SAT9 Norm book
   - Check the consistency of cut-off scores for DSTP-I and cut-score ranges for DSTP-II with the categories of proficiency levels
   - Check the application of the rules in writing
       - 997 = 0 + 0; 999 = 0 + 999 or 999 + 999; Valid writing score = valid + valid or valid + 999 or valid + 0;
       - 0-score: off-topic or foreign language
       - 997 = valid + 0; 999 = valid + 999 or 999 + 999; Valid writing score = valid + valid
   - Check the accuracy of scoring using the answer keys with a random sample by grade and test (1 percent)
   - Check the individual total score by grade and test using the keys
   - Alert the unusual data and report to the Assessment Director
   - Calculate the state, district, and school statistical summaries for DSTP-I, including the summary for ILCs at the school-level
   - The same checking will be applied for DSTP-II
   - Check the consistency of the summary data for DSTP-I between DDOE and HEM
   - The checking focuses on the n-counts and means and standard deviations on both SAT9 and SBS
   - The percentage of students at each proficiency categories on the DSTP-I and DSTP-II by grade and test at the state, district, and school levels
   - In house-checking between IT and Assessment summary data
   - Checking between DDOE and HEM summary data for the following files

5. Check Summary Files with HEM’s
   - Grades 2 to 10 by grade, test, state, district, and school: SAT9 – N-count, PR, NCE (mean, SD), scaled score, raw score; SBS – N-count, mean, SD; PS – DSTP-1, DSTP-2
   - Grade 10 re-testing by test, state, district, and school: SAT9 – N-count, raw score; SBS – N-count, mean, SD; PS – DSTP-1
   - Mixed-level testing by special code and test: SAT9 – N-count, raw score; SBS – N-count, mean, SD; PS – DSTP-1
   - Out-of-level testing by special code (grade) and test: SAT9 – N-count, raw score; SBS – N-count, mean, SD; PS – DSTP-1
   - Braille – a list by grade, test, scores, PS
   - List of students who took the wrong form/test
APPENDIX B
Accountability Score Quality Control Procedures

Delaware

1. Participation Target Status
   a. Review number of cells that are eligible for participation review. Cells with <40 students are excluded. [screen 2]. Six cells for each subject are eligible.
   b. Check calculations for appropriate cells, cells with large font. 95 percent participation of eligible students is required.
   c. Confirm that data agrees with the status as listed at the top of the screen, meets or does not meet participation.
   d. Confirm that [screen 3] information agrees with this screen, [screen 2].

2. Performance Target Status
   a. Confirm that summary screen [screen 1] information agrees with this screen, [screen 3].
   b. On [screen 3], review the number of cells with 40 or more students (note: on occasion this set of cells may not coincide with the performance cells on [screen 4] because performance data are weighted students). In the example school, 12 cells have the required students and 4 of these cells met or exceeded the target, two in ELA and two in Math. Therefore, annual performance was not attained.
   c. The performance details are shown on [screen 5]. Check that the appropriate cells are listed in the large font.
   d. Check that cells met target or did not meet target with and/or without confidence interval (CI) test. If CI was needed an “R” is shown.
   e. In the example Black, Hispanic, ELL and ED cells were below target in each subject.
   f. Check a percentage calculation on a few cells.

3. Safe Harbor Status
   a. Access safe harbor status screen [screen 6]. Check heading for agreement with [screen 1].
   b. Check to confirm that cells that did not meet AYP target [screen 5] are entered on this screen with the appropriate symbol, S, H, N or X, as defined on [screen 6].
   c. Check that the total number of cells not meeting safe harbor agrees with [screen 4]; in the example school, eight did not meet the target.

4. Other Indicator Status
   a. High school example: check graduation rate and progress on [screen 7]. Prior-year and previous year data are shown. “All Students” line is used; example school was above target.
   b. Elementary and middle schools: [screen 7a]. Improvement in average scale score is considered for students scoring below the standard.
      i. Did the average scale score of students below the standard (PL1 and PL2) increase?
      If not, did the percentage of students at Performance Level 1 (PL1) decrease?
      ii. If neither the average scale score improved nor the percentage in PL1 was reduced, a CI will be used to check both of the above, in the order as listed. The symbols listed on [screen 7a] document the results of these checks.
      iii. If none of the four tests were met, the Other Indicator was not met.
      iv. Check header on [screen 7a] with [screen 1].
Delaware (cont.)

5. AYP Status
   a. [Screen 8], check that cells are “meeting” (-) or “not meeting” (N) and that safe harbor coding agrees with [screen 6]. In the example, four cells attained the proficiency target. Check [screen 4] and two other cells met the target through safe harbor with a CI (H).
   b. The six “N”s indicate the cells that did not attain the target
   c. Check heading for agreement with [screen 1].

6. State Progress Status
   a. [Screen 9] summarizes the computations from [screen 10].
   b. Check that the information on the heading (composite score and progress) agrees with [screen 10].
   c. Check the full computation for one school to ascertain that the math was correctly executed.
   d. Look up prior year information for a school to ascertain that it had been accurately transferred to [screen 10].
   e. Check that rating agrees with table information on [screen 9].

7. School Rating
   a. Check that heading on screen 11 agrees with cover screen [screen 1]. In example the school is “Below” in both AYP and State progress.
   b. Check that the “Below-Below” pattern agrees with the rating scheme (Academic Progress or Academic Watch).
   c. Since the example is rated as Academic Watch with “Under School Improvement”, the school must have had prior year problems. This must be confirmed by screen 13.

8. School Improvement Status
   a. Screen 12 shows the historical patterns that lead to a school being Under School Improvement.
   b. Check that one of these patterns applies to the school. Screen 13 shows three-year the history for the school.
   c. Missing AYP in the same subject for two consecutive years classifies the school as Under School Improvement. Similar issues apply to Participation and to Other Indicators. The example school did not attain AYP for three years in either subject.
### Quality Assurance Diagnostic Matrix

<table>
<thead>
<tr>
<th>Stage</th>
<th>System Components</th>
<th>IT</th>
<th>Assessment</th>
<th>Accountability</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Efficient-Transparent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Systemic-Dynamic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Standardized-Validated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Formal-Compartmental</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Mixed Formality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monitored-Documented</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Planned-Undocumented</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Informal-Random</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Not Addressed</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Integrates policy changes
- Validates program membership
- Trains data providers
- Examines field specs, business rule narratives
- Determines data quality needs
- Trains data users
- Manages quality assurances
- Examines end-users feedback
- Conducts data audits
- Validates score production
- Operationalizes policy
- Supports accurate interpretations
- Promotes judgment credibility
- Aligns system components
- Evaluates behavioral changes

**APPENDIX C**

Quality Assurance Diagnostic Matrix (QADM)
## APPENDIX D
### QADM ACTION PLAN TEMPLATE

<table>
<thead>
<tr>
<th>Goal(s)-Subordinate Tasks</th>
<th>System Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>□ Information Management</td>
</tr>
<tr>
<td></td>
<td>□ Assessment</td>
</tr>
<tr>
<td></td>
<td>□ Accountability</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Process of Focus</th>
<th>Current Stage</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Improvement Activities</th>
<th>Program Manager</th>
<th>Milestones Dates</th>
<th>Resource Allocation</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time</td>
<td>Money</td>
</tr>
</tbody>
</table>

Validity Threats: Detection and Control Practices for State and Local Education Officials
## APPENDIX E
### West Virginia Five Year Plan for Accountability Validity

<table>
<thead>
<tr>
<th>Issue</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the cross-issue activities for accountability system development?</td>
<td>West Virginia will work with their TAC team to examine validity in the state accountability system.</td>
<td>West Virginia will work with their TAC team to examine validity in the state accountability system.</td>
<td>West Virginia will work with their TAC team to examine validity in the state accountability system.</td>
<td>West Virginia will work with their TAC team to examine validity in the state accountability system.</td>
<td>West Virginia will work with their TAC team to examine validity in the state accountability system.</td>
</tr>
<tr>
<td></td>
<td>The West Virginia Committee of Practitioners will review the policies and procedures of the accountability system.</td>
<td>The West Virginia Committee of Practitioners will review the policies and procedures of the accountability system.</td>
<td>The West Virginia Committee of Practitioners will review the policies and procedures of the accountability system.</td>
<td>The West Virginia Committee of Practitioners will review the policies and procedures of the accountability system.</td>
<td>The West Virginia Committee of Practitioners will review the policies and procedures of the accountability system.</td>
</tr>
<tr>
<td></td>
<td>The School Improvement Management Team will review the policies and procedures of the accountability system.</td>
<td>The School Improvement Management Team will review the policies and procedures of the accountability system.</td>
<td>The School Improvement Management Team will review the policies and procedures of the accountability system.</td>
<td>The School Improvement Management Team will review the policies and procedures of the accountability system.</td>
<td>The School Improvement Management Team will review the policies and procedures of the accountability system.</td>
</tr>
<tr>
<td></td>
<td>The WV State Board of Education will review and approve the policies and procedures of the accountability system pursuant to the State Board Strategic Plan.</td>
<td>The WV State Board of Education will review and approve the policies and procedures of the accountability system pursuant to the State Board Strategic Plan.</td>
<td>The WV State Board of Education will review and approve the policies and procedures of the accountability system pursuant to the State Board Strategic Plan.</td>
<td>The WV State Board of Education will review and approve the policies and procedures of the accountability system pursuant to the State Board Strategic Plan.</td>
<td>The WV State Board of Education will review and approve the policies and procedures of the accountability system pursuant to the State Board Strategic Plan.</td>
</tr>
<tr>
<td></td>
<td>Develop a survey of teachers and administrators —stratified sample of schools/districts that have been identified and have not been identified for school/district improvement.</td>
<td>Continue the accountability survey across years, resample, and resurvey.</td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
</tr>
<tr>
<td></td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
<td>Collect data to compare a growth model to the current status model. Use data collected to propose (1) a growth model, (2) a proficiency index model or (3) other changes to the West Virginia Accountability Plan.</td>
</tr>
</tbody>
</table>

**Validity Threats:** Detection and Control Practices for State and Local Education Officials
<table>
<thead>
<tr>
<th>How valid and reliable are the AYP decisions?</th>
<th>Review the validity of the AYP components: attendance rate indicator, graduation rate indicator, minimum &quot;N&quot; for accountability rate, and full academic year.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review the data definitions, procedures, and collections related to AYP decisions.</td>
<td>Review data accuracy.</td>
</tr>
<tr>
<td>Review changes resulting from the OEPA appeals process.</td>
<td>Review the results of the assessment appeals process.</td>
</tr>
<tr>
<td>Use the CAS triage tool for determining level technical assistance provided to schools/districts.</td>
<td>Document the Webb alignment study for the WESTEST:</td>
</tr>
<tr>
<td>• internal</td>
<td>• external</td>
</tr>
</tbody>
</table>

Document and review WESTEST bias reviews:
<table>
<thead>
<tr>
<th>How effective are the consequences in effecting the intended changes?</th>
<th>Use impact data to study the intended and unintended consequences of the accountability system.</th>
<th>Use impact data to study the intended and unintended consequences of the accountability system.</th>
<th>Use impact data to study the intended and unintended consequences of the accountability system.</th>
<th>Use impact data to study the intended and unintended consequences of the accountability system.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determine if subgroup performance is influenced by applied consequences.</td>
<td>Determine if subgroup performance is influenced by applied consequences.</td>
<td>Determine if subgroup performance is influenced by applied consequences.</td>
<td>Determine if subgroup performance is influenced by applied consequences.</td>
<td>Determine if subgroup performance is influenced by applied consequences.</td>
</tr>
<tr>
<td>Examine the impact of school choice and supplemental services on school improvement.</td>
<td>Examine the impact of school choice and supplemental services on school improvement.</td>
<td>Examine the impact of school choice and supplemental services on school improvement.</td>
<td>Examine the impact of school choice and supplemental services on school improvement.</td>
<td>Examine the impact of school choice and supplemental services on school improvement.</td>
</tr>
<tr>
<td>Compare NAEP, ACT Plan, ACT Explore, and WESTEST science and social studies to RLA and Math to determine if there is generalized improvement.</td>
<td>Compare NAEP, ACT Plan, ACT Explore, and WESTEST science and social studies to RLA and Math to determine if there is generalized improvement.</td>
<td>Compare NAEP, ACT Plan, ACT Explore, and WESTEST science and social studies to RLA and Math to determine if there is generalized improvement.</td>
<td>Compare NAEP, ACT Plan, ACT Explore, and WESTEST science and social studies to RLA and Math to determine if there is generalized improvement.</td>
<td>Compare NAEP, ACT Plan, ACT Explore, and WESTEST science and social studies to RLA and Math to determine if there is generalized improvement.</td>
</tr>
<tr>
<td>Review Title I, Special Education, OEPA monitoring process to determine school/district improvement.</td>
<td>Review Title I, Special Education, OEPA monitoring process to determine school/district improvement.</td>
<td>Review Title I, Special Education, OEPA monitoring process to determine school/district improvement.</td>
<td>Review Title I, Special Education, OEPA monitoring process to determine school/district improvement.</td>
<td>Review Title I, Special Education, OEPA monitoring process to determine school/district improvement.</td>
</tr>
</tbody>
</table>