

The Optimal Reference Guide:

The Data Quality Manual

Data Quality Series – Part II

Extraordinary insight™ into today's education information topics

By Glynn D. Ligon



ESP Solutions Group

Table of Contents

Introduction	3
Steps for Achieving Data Quality.....	3
Data Quality 101	4
The “don’ts” to avoid messing up your data.....	4
Software Vendors.....	7
The 80/20 Rule of Data Quality.....	8
Checklist for Sensing the Quality of Data.....	9
The Data Quality Rating Scale	12
The Four Great Truths about Data Quality	16
Steps for Ensuring Data Quality.....	16
A Final Note about Error.....	16
Conclusion	16



Introduction

Let's get to work and improve data quality.

Data quality matters now.

Data quality is an official buzz word.

Data quality steps are known now.

Data quality is for everyone.

Data quality is detectable.

Data quality saves money.

Data quality relieves stress.

By assimilating the conventional wisdom about data quality with the real school experiences of our ESP professionals, we have been able to create a tutorial on the practices that cause bad data and the processes that ensure quality data.

Steps for Achieving Data Quality

The authors assisted the U.S. Department of Education in the development of a set of data quality standards for program data. A training package was developed from those standards and sessions were conducted with program office staff. We took those relatively high-level standards and created a step-by-step process for managing the quality of data across an entire annual cycle.

Data Quality 101

The “don’ts” to avoid messing up your data

Never, ever create a reporting format that allows for:

- leading or trailing zeroes
- repeated numbers or letters in an identifier or code
- mixing numbers and letters in an identifier or code unless 0, O, I, 1, l, and all other confused characters are left unused

 **ESP Insight**
These “don’ts” were collected across dozens of data management projects.

The most frequent and insidious errors that plague an information system:

DO NOT:

1. Make notes in data fields.

First Name Field:

“Mandy (but mother says she prefers to be called “Pookey”)”

2. Copy and paste from one file (format) to another.

Pat	M	Johnson	Jr
Johnson, Pat M, Jr			

3. Be lackadaisical when the requirements are precise.

Patrick	M.	Johnson	Jr.
Pat		Johnson	

4. Add codes to be more specific.

1 = Graduate

2 = Transfer

3 = Retainee

U = Unknown

M = Sent to Marie for Coding

5. Make the data your own.

Phone Number Field:

"555-555-5678 except on Tue then 656-555-5555"

6. Give everyone the same value just to fill the field.

SSN Field:
"111-11-1111"

7. Submit split or duplicate records.

Student Name	Birthday	Test Score	Course Grade	Absences
Pat Johnson	09111999	98	A	3
Pat Johnson	09111999	98	A	3
Kelly Smith	12251999	79		8
Kelly Smith	12251999		B	8

8. Ask for forgiveness rather than permission.

"Oh, hello, yeah, I think I may have accidentally left all the Title 1 codes off my file. I'm really sorry. Can you ever forgive me?"

9. Argue with official names, spelling, or capitalization.

District Name Field:
"Colorado Springs"
(Official Name: El Paso County District 11)

10. Be right when the world is wrong.

Street Name Field:
"Arroyo Seco"
Arroyo Seca is the official name.

11. No matter how dumb they act, don't say students were born yesterday.

Birth Date Field:
"April 11, 2008"

12. Be creative to get double use from the data.

Course Field:
"Lunch A"

13. Be better when the software is good enough.

Gender Field:
"Female"

(Valid Code = F)

14. Keep doing things the way you did before the new software was installed.

“My Excel spreadsheet is really the official record for my students.”

15. Call a friend at the district office or SEA and ask for her/him to correct your data.

“Hi Coleen, would you be a dear and just change those LEP codes for me again this year?”

16. Copy and paste without being extra careful.

Grade	Gender	First Name	Last Name
7	M	Freddy	Hanson
8	M	Sandra	Hernandez
7	M	Charlotte	Webster
6	M	John	Johnson
6	M	Michelle	Michelle
7	M	Juan	Paredes
7	M	Janelle	Smith
8	M	Herbert	White
8	M	Snoop	Perro

17. Think of data quality as an as-of-date requirement (wait to get everything right on the reporting date).

18. Pass data entry on to someone who doesn't know the rules or can't follow them.

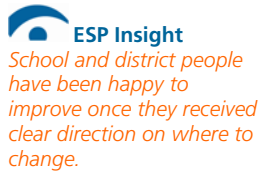
“Our student aide will enter everything. Huh? FERPA? Training?”

Software Vendors

Software vendors are your partners. Better yet, software vendors are your “employees.” They need to make you look good. You must insist they follow the rules. Of course, this means that the people paying the vendors must insist. That may be the SEA or the LEA, or at times an individual school or program.

When we began the first statewide data collections using SIF in Wyoming, the “SIF certified” agents for student information systems (SIS) sent data to State Report Manager (SRM), ESP’s product for collecting and verifying data for the Wyoming Department of Education (WDE), using whatever codes they found in each district’s SIS. SRM’s business rules flagged them as fatal errors. This began a nationwide effort to accomplish two objectives. First, SIFA had to enhance their certification process to require that agents follow the complete standard including use of approved codes. Second, the SIS vendors had to enhance their agents to crosswalk or accept only approved codes. If the line had not been drawn in the sand at that point, the WDE staff would have continued to fix each district’s submission file before certifying the collection to be complete and ready for use.

 **ESP Insight**
*Software vendors are your
partners. Better yet,
software vendors are you
“employees.”*



The 80/20 Rule of Data Quality

You can either put in 80% of the effort cleaning up the data every year—or only 20% of the effort up front to establish clear rules and insist they be followed. Yes, that 20% is a lot of effort up front. Standard operating procedure is that work is done just good enough at each step because someone later on will clean things up if it's really that important. That's unacceptable. The 80/20 rule has been changed in Wyoming and other states using SRM as a gatekeeper for data quality to the 20/2 rule. That's 20% of the effort is invested up front to ensure all business rules are met and only 2% of the effort from then on to handle outliers.

The greatest benefit has accrued to the local schools and districts. Using the specific, user-friendly edit reports that SRM provides as their trial data are tested, they have improved their processes to avoid entering or perpetuating many of the data problems that were inherent in the legacy systems. School and district people have been happy to improve once they received clear direction on where to change.

Process Flow of Reported Data:

- **Declaring by the original source of the data (parent)**
- **Entering by the collector**
- **Compiling for reporting**
- **Sending**
- **Receiving**
- **Mapping**
- **Importing**
- **Accessing**
- **Analyzing**
- **Formatting**
- **Labeling**
- **Explaining**
- **Interpreting**
- **Using**

Checklist for Sensing the Quality of Data

Sometimes the best way to determine the likelihood of quality data is for a human being to stare at the numbers and see if they make sense. Read *Blink: The Power of Thinking Without Thinking*, 2007, Malcolm Gladwell, to see how much of an expert you probably are when it comes to your own statistics.

From decades of proofing data reports, Figure 1 summarizes some ideas for checking the data for possible errors. Steps 1 through 12 are somewhat in order of their sophistication, but number 13 sums up the lesson from *Blink*—What’s your gut reaction?

 **ESP Insight**
Sometimes the best way to determine the likelihood of quality data is for a human being to stare at the numbers and see if they make sense.

Figure 1 Steps for Validating Data

Step	Description	Example
1. Your Best Guess	Write down your best guess of what the statistic should be. How close to your prediction is the reported statistic?	From all you've read, you know that reported dropout rates range considerably, but you expect the local rate to be about 3% a year. The preliminary rate sent to you from MIS is .35%. (Correcting an errant decimal made the rate 3.5%. That's reasonable.)
2. Prior Statistic	Find a previously reported statistic, preferably several across reporting times. How close to prior trends is the reported statistic?	The prior four years' dropout rates have been 6.7%, 5.4%, 3.8%, and 3.4%. So, 3.5% looks reasonable.
3. Another Entity	Find statistics for similar entities (e.g., other schools, states, programs). Write down your best guess of how they should compare. How do the statistics actually compare?	The statewide dropout rate for the prior year was 4.1%. The neighboring district reported 2.9%. Because your district is roughly between the two in demographics, you guess that your local rate should also be between theirs. 3.5% looks logical.

Step	Description	Example
4. Simple Math	Do some simple math with the statistic. Do the results make sense?	The technology report states that students average 2 hours a week on computers. You know the number of hours in a school day, the number of students, and the number of computers. Your simple calculations show that if every computer had a student on it every minute of the day, the average could only be 2 hours a week. Such efficient scheduling is impossible.
5. Calculate Counts	If the statistic is a percent, proportion, or ratio, calculate an actual count. Does this count make sense?	The report draft showed 12% of the students enrolled in AP English at the high school. That would be about 200 students. With only one AP English teacher, this doesn't seem right.
6. Calculate Percents	If the statistic is a count, calculate a percent, proportion, or ratio. Does this calculation make sense?	The report showed 267 students eligible for a free lunch. That would be about 18% of the high school students. The high school must have at least 35% because it is one of your Title I schools.
7. Know the Source	Who is reporting the statistic? Are they the right person to do so? Are they the original source? Do you trust them?	The district's music coordinator writes that 67% of college scholarship recipients were music students when in middle and high school. No source for the statistic is cited. You check and find that 67% of parents responding to a band booster survey said their child would receive some financial aid.
8. Independent Verification	Was the statistic independently verified?	The superintendent states that 82% of the district's students passed the statewide math exam. The statistic is also reported by the state education agency and was calculated by the vendor for the assessment program.



Step	Description	Example
9. Graph Proportions	If there is a graph, are the scales and proportions appropriate?	A graph shows a dramatic increase in the number of students taking algebra. The y axis begins at zero and goes above the highest value shown. The ratio of the y to x axes is about 3 to 4. Everything appears to be done just like the text books suggest. So the impressive look of the graph is appropriate.
10. Details and Documentation	Are definitions, measures, limitations, samples, and other information provided for judging the validity of the statistic?	The evaluation that reported the algebra enrollments is accompanied by a technical report with the details.
11. Definitions and Periodicities	Do comparisons or changes reported use the same data points, definitions, periodicities, etc.?	Some problems are evident with the algebra enrollments. The current year is based upon beginning of the semester enrollment, but past years are counts of students earning credit. Past years include summer school, but the current year's summer is still in progress.
12. Stakes	What's at stake? How might the stakes have influenced the reporting of the statistic? How would competing perspectives have interpreted the statistic?	The high school is applying for a grant and must include achievement gains. The gains are impressive, but a change in school boundaries moved a large number of higher achieving students into the school last year. No adjustment for these students was made to verify that gains were made by the continuously enrolled students.
13. Gut Reaction	What's your gut reaction?	The district reports that dropouts have declined by 75% over the past five years. You haven't noticed great changes, new programs, or any other intervention that could make such a huge difference. Reaction: You doubt this one.




The Data Quality Rating Scale

Use this to determine how good your data are.

Consumer Reports would want us to provide a rating system for data quality, so here's one (Figure 2). Using the criteria of validity, accuracy, lateness, usefulness, and expense, an information source can be rated on this five-level scale. Try an area of data you are familiar with and apply the ratings. When I did this for the information systems I used to manage, the surprising winner was food service data. The loser? Discipline data. Make that undisciplined data.



Figure 2 Data Quality Rating Scale




Information Source:				Source Type:	
	Validity	Accuracy	Lateness	Usefulness	Expense
 High Quality	There is a clear match between the data and the intended or primary use of the data. Appropriate comparisons can be made. Appropriate conclusions can be made.	Data are accurate and complete. Data standards are clear and were followed.	The most recent data are provided. The time period of the data match the use and intent of the data.	Data are presented completely and clearly for ease of use. Access to the data for use is easy.	No charge is made for access or use.
 Reduced Quality	A relationship between the data and the intended or primary use of the data is assumed or is logical, but may not be well documented or proven.	Data standards are documented. Compliance is assumed to be reasonable. Limitations are described.	Data are recent enough to suggest reasonable applicability for use and intent.	Data are presented well for use. Access requires some effort but is available.	Copies or access is free, but some charges apply.

Information Source:				Source Type:	
	Validity	Accuracy	Lateness	Usefulness	Expense
 Low Quality	The connection between the data and the use of the data is weak or nonexistent.	Data standards are weak or nonexistent. Poor controls are in place to ensure compliance.	Aged data may not be appropriate for making decisions about current issues.	Data are poorly presented or explained. Access is cumbersome and limits use.	A charge applies for access or use.
 Poor Quality	The connection between the data and the use of the data is misrepresented or misleading.	Incorrect data, substantial missing data, or other problems are evident.	Data are too old to be useful.	Data are uninterpretable or inaccessible.	A substantial charge applies for access or use compared to similar sources of information.
 Unknown Quality	How well the data and the use of the data match is not known or not described.	Accuracy of the data is unknown or not documented.	The periodicity is unknown. The appropriateness of the data is unknown because of the lateness of them.	Unknown.	Unknown.

Each of the rating components needs to be further detailed to ensure comparable ratings across raters. Accuracy is presented in Figure 3 as an example.

Figure 3 Accuracy Scale

Accuracy		The Data are Rated at the Level in Which ALL Conditions are Satisfied.			
 High Quality	Data are accurate and complete. Data standards are clear and were followed.	81-85: A. Missing data are not well documented and impact use minimally. B. Data are certified by providers as accurate; problems are documented. C. Data standards and specifications are published and readily available to providers.	86-90: A. Missing data are well documented and impact use minimally. B. All data are certified by providers as accurate. C. Data standards and specifications are published and providers certify their compliance.	91-95: A. Missing data are well documented and do not impact use. B. All data have been verified as accurate by the collecting agency. C. Data standards and specifications are published and data are checked for compliance.	96-100: A. No data are missing. B. All data have been certified as accurate through audit or review. C. Data standards and specifications are published and data are in compliance.
 Reduced Quality	Data standards are documented. Compliance is assumed to be reasonable. Limitations are described.	61-65: A. Missing data limit use in at least one key area. B. Data problems are evident and limit use. C. Data standards and specifications are not relied upon.	66-70: A. Missing data limit use. B. Data problems are evident and may limit use. C. Data standards and specifications are not relied upon consistently.	71-75: A. Missing data are not documented and use is impacted. B. Data problems not documented and may limit use. C. Data standards and specifications do not provide adequate guidance to data providers.	76-80: A. Missing data are not well documented and use is impacted. B. Data problems are not fully documented and may limit use. C. Data standards and specifications are partially complete or in need of updating.

 Low Quality	Data standards are weak or nonexistent. Poor controls are in place to ensure compliance.	41-45: A. Most key data are missing. B. Data problems are pervasive and prevent use. C. Data standards and specifications are not available.	46-50: A. Substantial, key data are missing. B. Data problems are pervasive and prevent most use. C. Data standards and specifications are not available.	51-55: A. Missing data are prevalent enough to substantially limit use. B. Data problems are pervasive and substantially limit use. C. Data standards and specifications are not available.	56-60: A. Missing data are prevalent enough to require caution in use. B. Data problems are evident and substantially limit use. C. Data standards and specifications are not relied upon.
 Poor Quality	Incorrect data, substantial missing data, or other problems are evident.	0-10: A. Most data are missing. B. All data exhibit major problems. C. Data standards and specifications are not available.	11-20: A. Most data are missing. B. All data exhibit problems. C. Data standards and specifications are not available.	21-30: A. Most data are missing. B. Data problems are universal. C. Data standards and specifications are not available.	31-40: A. Most data are missing. B. Data problems are substantial. C. Data standards and specifications are not available.
 Unknown Quality	Accuracy of the data is unknown or not documented.				

The Four Great Truths about Data Quality

Data quality is highest when...

1. The data providers know what's expected.
2. The data providers use the data themselves for their own work.
3. Everyone, everywhere checks the data.
4. The data are available and used.

Part 1 of the Data Quality Series, *The Data Quality Imperative*, identified these four truths about data quality. They guided the design of the steps outlined in Attachment A: Data Quality, Best Practices for Local, State, and Federal Education Agencies.

Steps for Ensuring Data Quality

All the above is well and good—if not great in places. However, for those professionals on the line, designing and managing programs and information systems, there needs to be a users guide for data quality. There is.

Attachment A takes the principles and insights from this paper and translates them into the day-to-day activities that must be followed to achieve the highest level on the hierarchy.

A Final Note about Error

The hierarchy and the detailed steps do not deal completely with some of the nitty-gritty issues of data quality that are usually fretted over by information systems managers and data providers. Many of these fall into the general category of error. Error can be mistakes that result in bad data. Those have been addressed already. Error can also be measurement error (such as the standard error of measurement for an assessment) that keeps us from ever being 100% confident in our data.

Measurement errors are those imprecisions that result from our inability to be absolutely perfect in our measurements. One is the reliability of an instrument, test, or performance task (illustrated by a test-retest difference). Measurement errors can also be “intentional” as occurs when we round numbers or put values in ranges rather than use a more precise value. In research and evaluation situations, sampling error introduces its own limits on the reliability of the data. Measurement error should be recognized and acknowledged when data make their way to the reporting end of their life cycle.

Conclusion

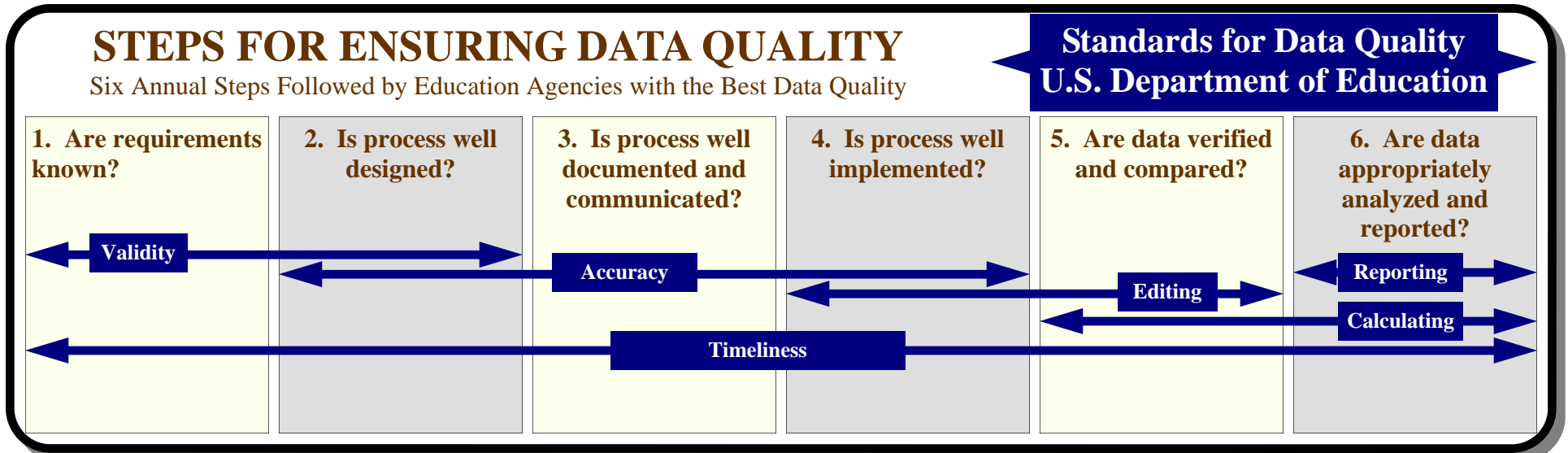
Data quality is achievable if we establish the rules and follow them—all of us.

Attachment A follows.

Data Quality

Best Practice for Local, State, and Federal Education Agencies

Glynn D. Ligon, Ph.D. & Barbara S. Clements, Ph.D.



Data Quality:

Data quality is more than accuracy and reliability. High levels of data quality are achieved when information is valid for the use to which it is applied and when decision makers have confidence in the data and rely upon them.

Information Systems Architecture:

The foundation for data quality begins with a formal information systems architecture (ISA). The ISA is the metadata, hardware, software, and network standards, policies, governance, and requirements by which all information systems are built and managed. See the D3M Framework as described in Our Vision for D3M at <http://www.espsg.com/resources.php>.

OVERVIEW

STEPS FOR ENSURING DATA QUALITY

Standards for Data Quality
U.S. Department of Education

1. Are requirements known?	2. Is process well designed?	3. Is process well documented and communicated?	4. Is process well implemented?	5. Are data verified and compared?	6. Are data appropriately analyzed and reported?
Validity		Accuracy		Editing	Reporting
		Timeliness			Calculating
<p>Compare policy, regulation, rules, and procedures with the instructions given to data providers, collection forms, and code in software applications. M</p> <p>Include data providers and data processors in decisions to establish what is feasible. M</p> <p>Follow an established change-management process. M P</p> <p>Comply with professional standards for data collection, analysis, and reporting. M E</p> <p>Ensure people at all levels are knowledgeable, certified, trained, and competent for the tasks for which they are responsible. M E</p>	<p>Review design by peers, agencies, and staff. M</p> <p>Precode all available data. Limit times data are entered. P</p> <p>Use most automated/validated level of data entry possible (e.g., selection from codes in an automated application vs. filling in fields). P</p> <p>Use random checks during production. P</p> <p>Automate verification of entries at the earliest levels (e.g., upon keying Vs. from printed audit report). P</p> <p>Run maintenance before all production. Verify off-hour maintenance and staff availability. P</p> <p>Ensure target dates are reasonable and clear. M</p>	<p>Provide training and certification for data providers. Train all new staff. M P</p> <p>Provide documentation for data providers and data processors. M P</p> <p>Provide immediate help for data providers. M P</p> <p>Ensure the physical and fiscal requirements are available (e.g., computer hardware, software, network, etc.) M P</p>	<p>Use checklists and sign-offs for key steps. P</p> <p>Run sample data and verify. P</p> <p>Ensure problems are identified, documented, corrected, and communicated back to the source of the problem or report. M P</p> <p>Conduct on-site reviews during the process. M P</p>	<p>Run audit reports for review by experts with knowledge of reasonableness. M E</p> <p>Verify all calculations and conditional/business rules. M P E</p> <p>Compare data to past runs, standards, or similar groups. M P E</p> <p>Check data exchanges, crosswalks, and translations for integrity. P</p>	<p>Fully disclose conditions affecting interpretation of the data. M P E</p> <p>Review data with providers and others with a stake in the results. M E</p> <p>Ensure analysis techniques meet the assumptions required for proper use. M E</p> <p>Present conclusions fairly within a context for interpretation. M E</p> <p>Publish technical reports or make available data files with detailed data for verification of analyses and statements. M E</p> <p>Protect the confidentiality rights of individuals (FERPA). M E</p>
<p>Persons Primarily Responsible for Data Quality During Each Step:</p> <p>M = Manager of the program; designer of the collections; collector of the data; data steward</p> <p>P = Computer programmer; designer of the processing; processor of the data</p> <p>E = Evaluator; analyst; report writer</p> <p>The provider of the data (e.g., school) is responsible for conscientiously following the prescribed process, reporting problems, and verifying the accuracy and completeness of all data submitted.</p>					

1. Are requirements known?

Validity

Timeliness

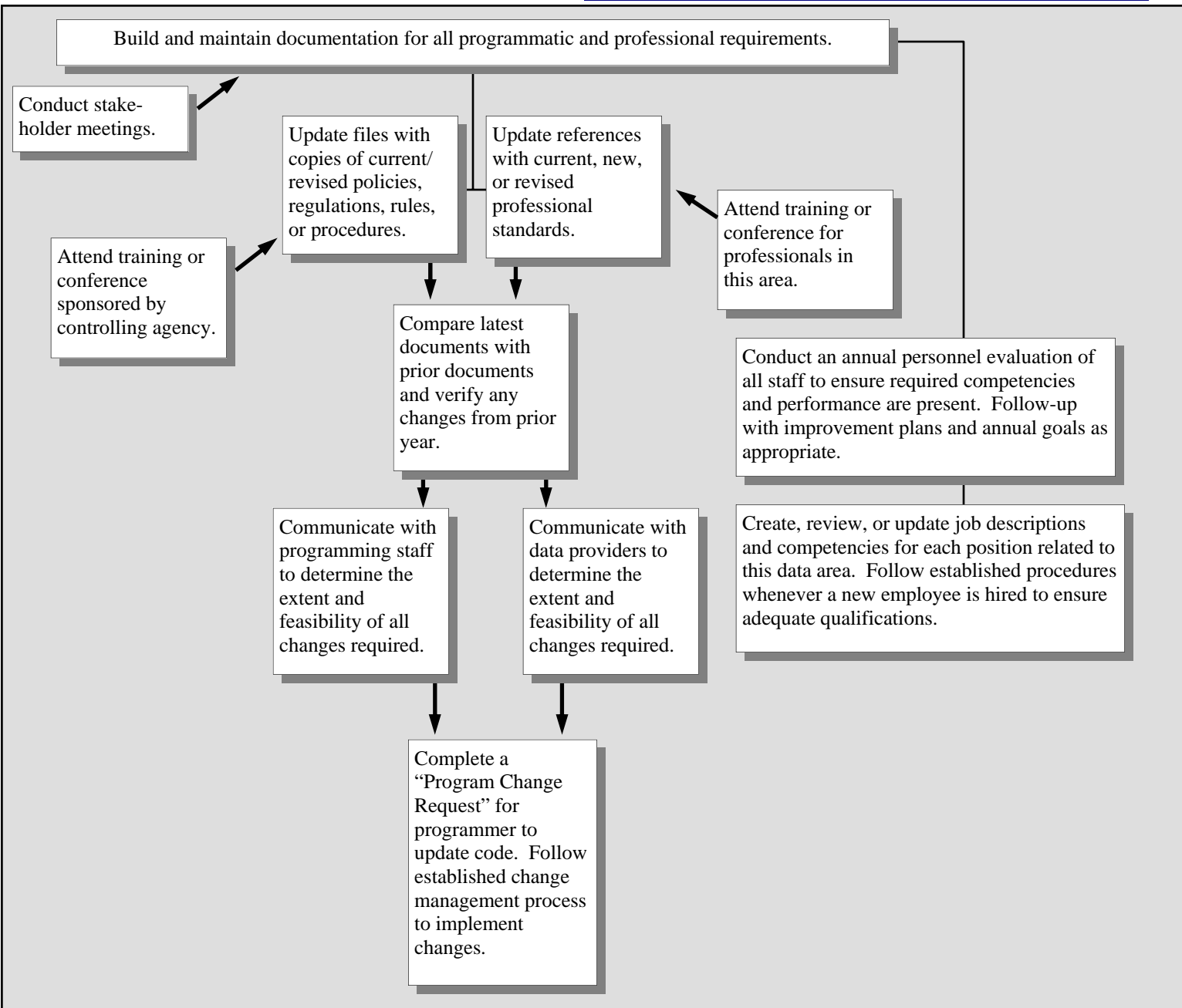
Compare policy, regulation, rules, and procedures with the instructions given to data providers, collection forms, and code in software applications. **M**

Include data providers and data processors in decisions to establish what is feasible. **M**

Follow an established change-management process. **M P**

Comply with professional standards for data collection, analysis, and reporting. **M E**

Ensure people at all levels are knowledgeable, certified, trained, competent, and energetic for the tasks for which they are responsible. **M E**



2. Is process well designed?

Validity

Accuracy

Timeliness

Review design by peers, agencies, and staff. **M**

Precode all available data. Limit times data are entered. **P**

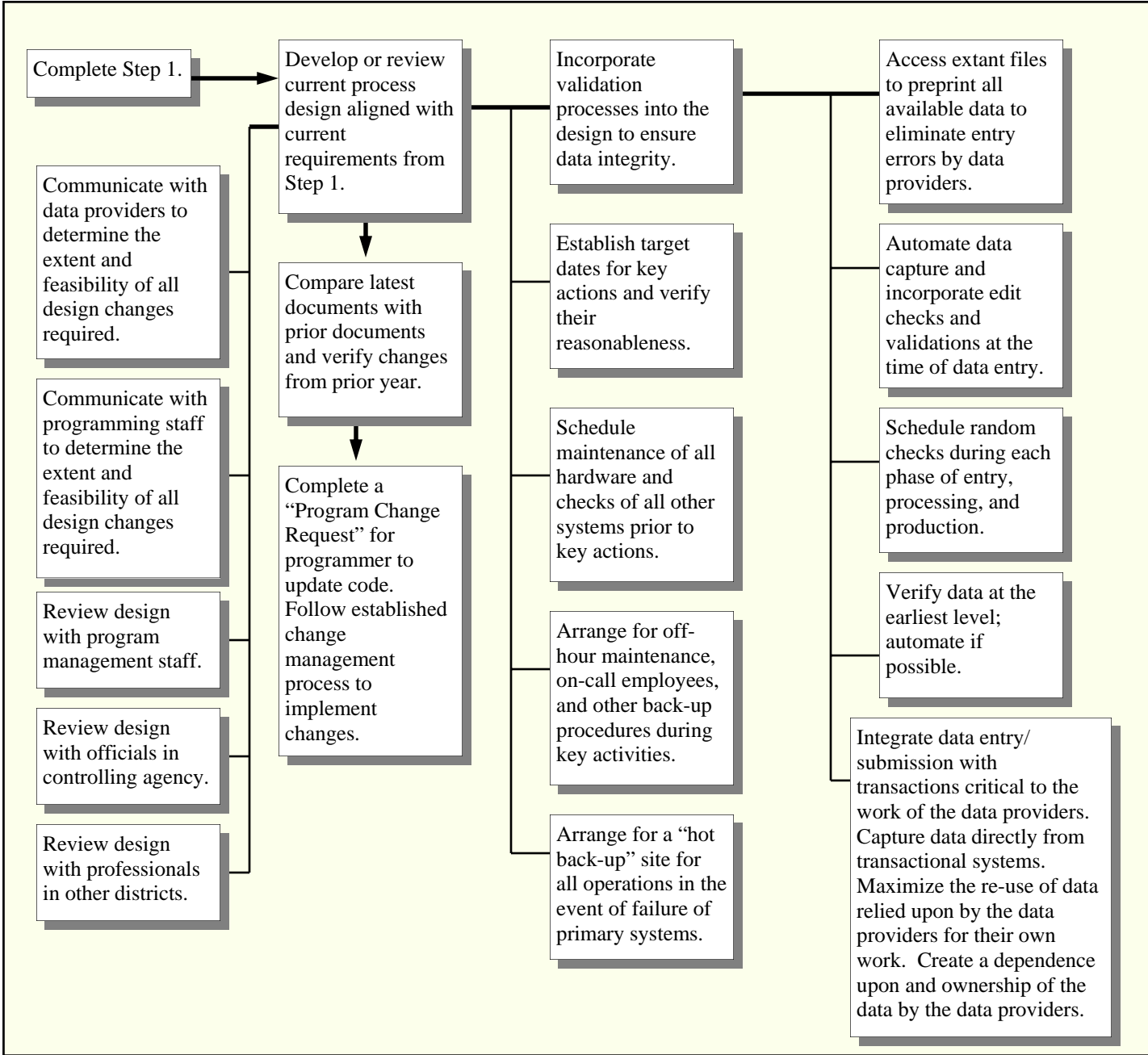
Use most automated/validated level of data entry possible (e.g., selection from codes in an automated application vs. filling in fields). **P**

Use random checks during production. **P**

Automate verification of entries at the earliest levels (e.g., upon keying Vs. from printed audit report). **P**

Run maintenance before all production. Verify off-hour maintenance and staff availability. **P**

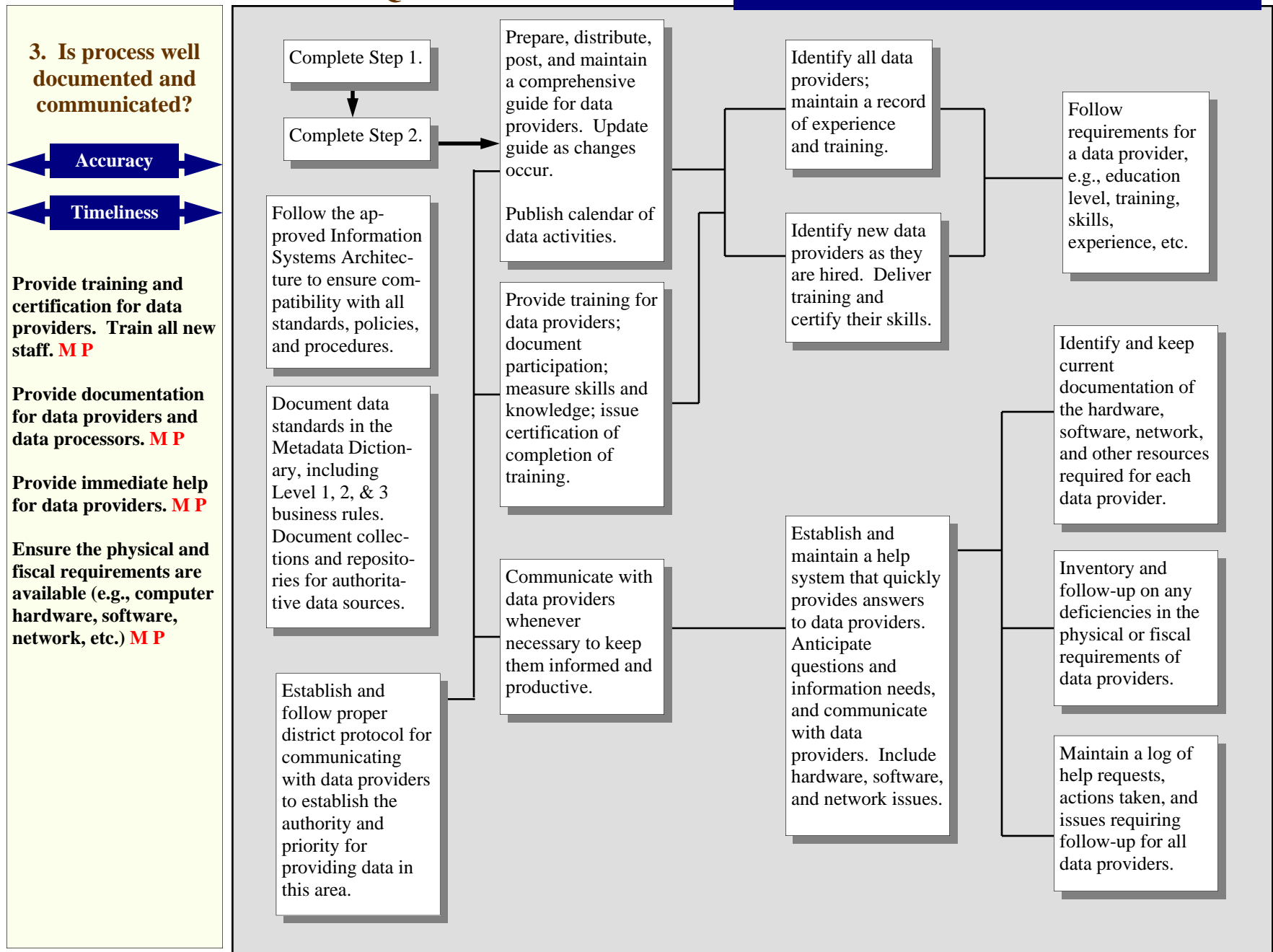
Ensure target dates are reasonable and clear. **M**



D3M Data Quality Process

Best Practice DATA QUALITY STEP 3

Standards for Data Quality U.S. Department of Education



4. Is process well implemented?

← Accuracy →

← Editing →

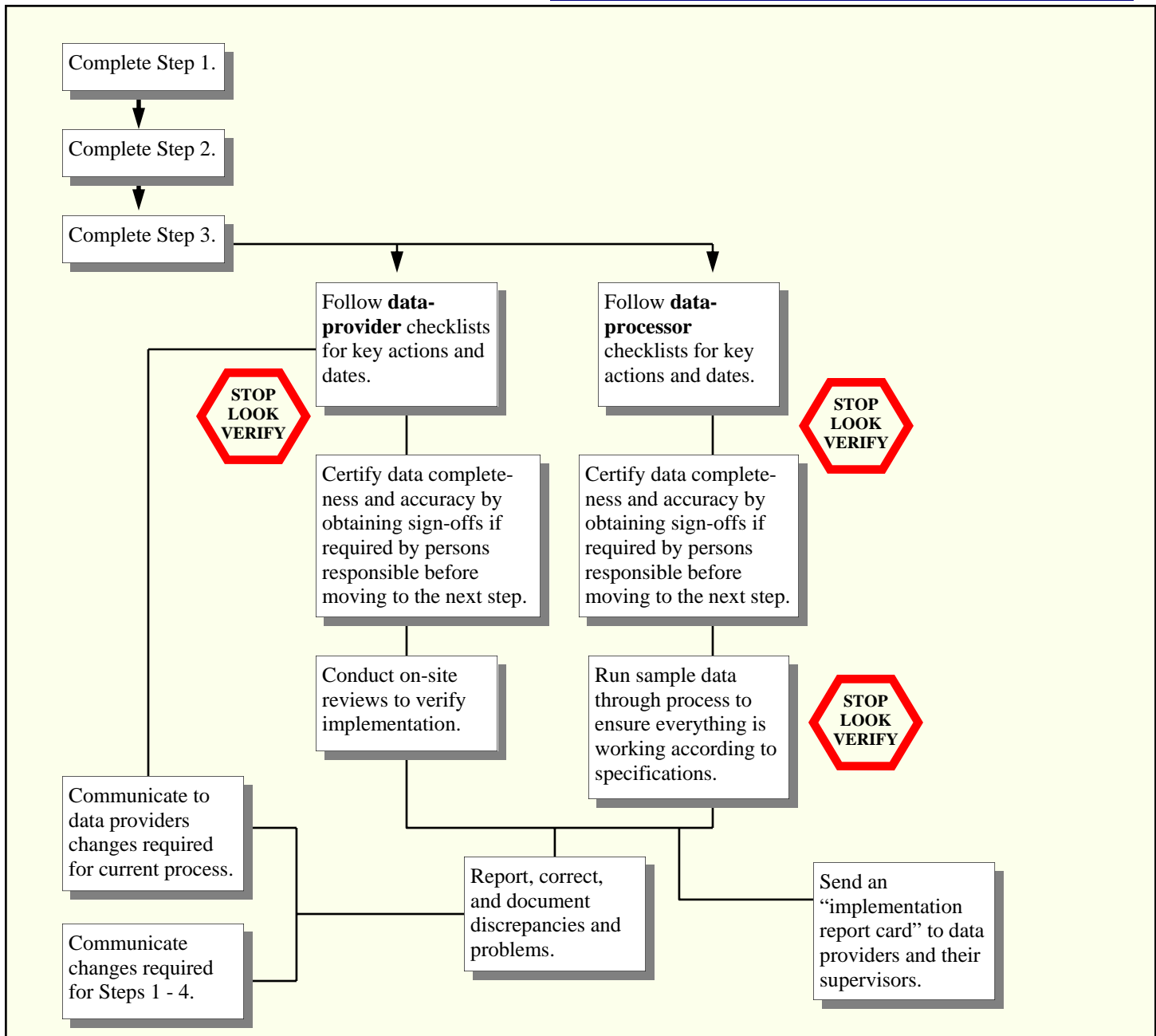
← Timeliness →

Use checklists and sign-offs for key steps. **P**

Run sample data and verify. **P**

Ensure problems are identified, documented, corrected, and communicated back to the source of the problem or report. **M P**

Conduct on-site reviews during the process. **M P**



5. Are data verified and compared?

◄ Editing ►

◄ Calculating ►

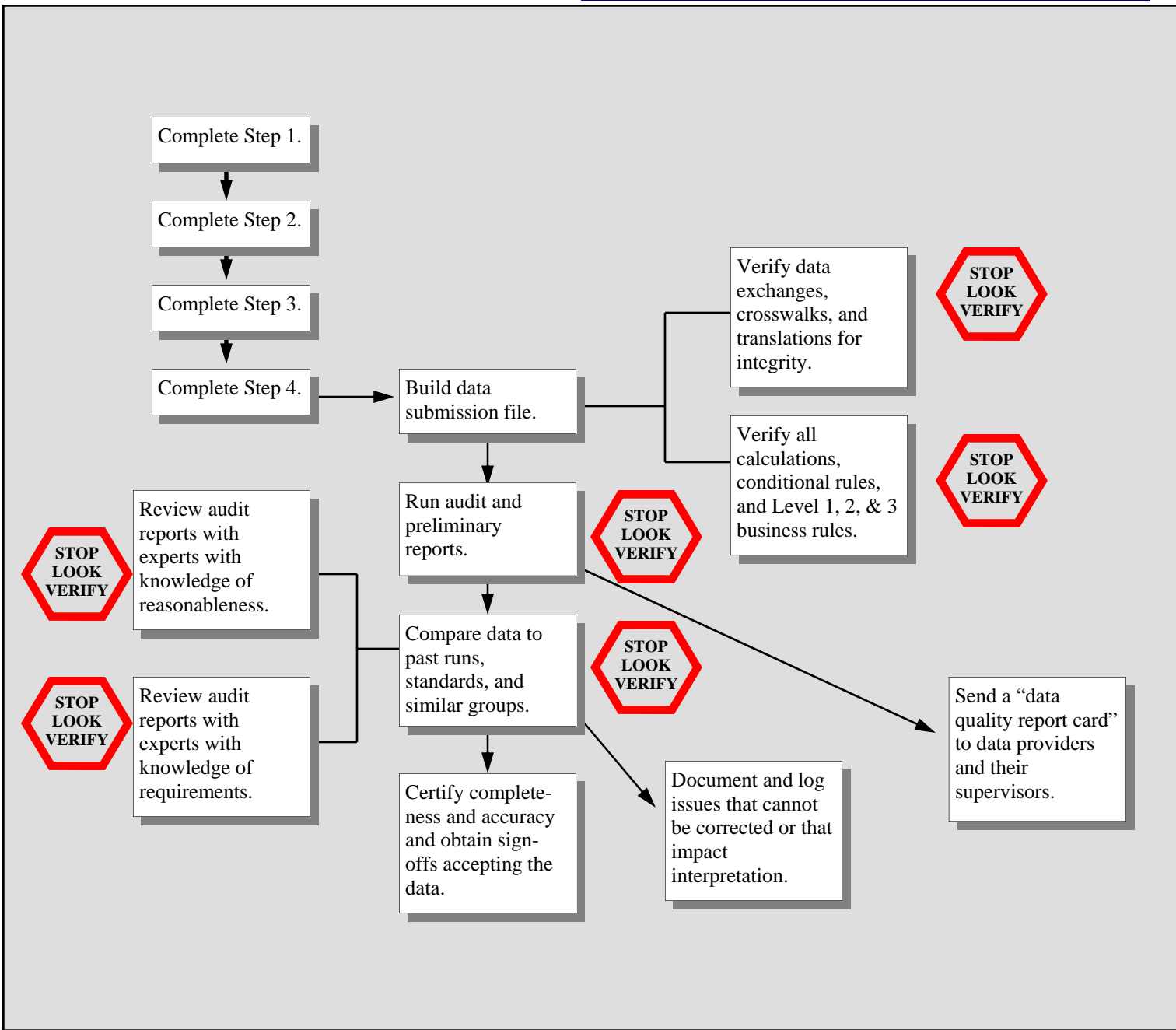
◄ Timeliness ►

Run audit reports for review by experts with knowledge of reasonableness. **ME**

Verify all calculations and conditional rules. **ME**

Compare data to past runs, standards, or similar groups. **MEPE**

Check data exchanges, crosswalks, and translations for integrity. **P**



6. Are data appropriately analyzed and reported?

Reporting

Calculating

Timeliness

Fully disclose conditions affecting interpretation of the data. **M P E**

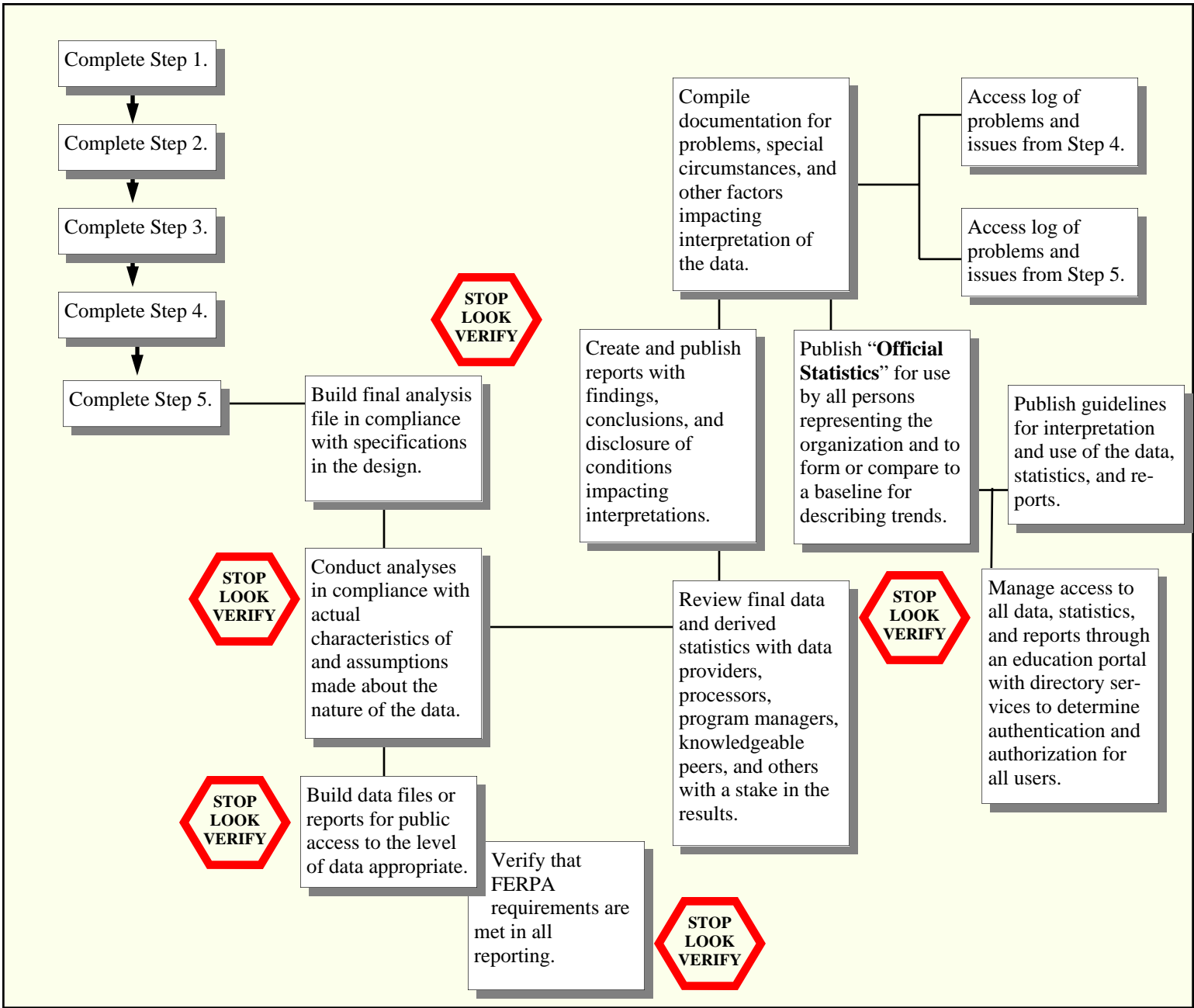
Review data with providers and others with a stake in the results. **M E**

Ensure analysis techniques meet the assumptions required for proper use. **M E**

Present conclusions fairly within a context for interpretation. **M E**

Publish technical reports or make available data files with detailed data for verification of analyses and statements. **M E**

Protect the confidentiality rights of individuals (FERPA). **M E**





About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight*™ into PK-12 education data systems and psychometrics. Our team is comprised of industry experts who pioneered the concept of “data driven decision making” and now help optimize the management of our clients’ state and local education agencies.

ESP personnel have advised school districts, all 52 state education agencies, and the U.S. Department of Education on the practice of K-12 school data management. We are regarded as leading experts in understanding the data and technology implications of the **No Child Left Behind Act (NCLB)**, **Education Data Exchange Network (EDEN)**, and the **Schools Interoperability Framework (SIF)**.

Dozens of education agencies have hired ESP to design and build their student record collection systems, federal reporting systems, student identifier systems, data dictionaries, evaluation/assessment programs, and data management/analysis systems.

To learn how ESP can give your agency *Extraordinary Insight* into your PK-12 education data, email info@espsg.com.

This document is part of *The Optimal Reference Guide Series*, designed to help education data decision makers analyze, manage, and share data in the 21st Century.

The Data Quality Manual, Data Quality Series – Part II. Copyright © 2008 by ESP Solutions Group. All rights reserved. No part of this paper shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher.



ESP Solutions Group

(512) 879-5300

www.espsolutionsgroup.com