Assuring Accurate Student Assessment Results:
Steps States and Their Contractors Can Take

Edward Roeber
Council of Chief State School Officers

January, 1998

Assuring Accurate Student Assessment Results: Steps States and Their Contractors Can Take

Edward Roeber

Introduction

Today, over forty-eight states operate large-scale student assessment programs. While the number of such programs that each state runs may range from one to six, the typical state is operating about three various assessment programs for different purposes. These range from low stakes needs assessment programs to ones with high stakes such as awarding or denying high school diplomas or rewarding or sanctioning schools for their performance on the assessments. Most of the assessment programs operating in the United States involve assessing all students in one or more grade levels, so that results are returned to the parents of the students tested and their teachers and school administrators. Testing in America affects many students, parents, teachers, school administrators and the public at large. The stakes for testing students accurately have increased with the increased visibility attached to the enterprise of testing.

Testing is also becoming more complex. While in the past it was typical to test all students at several grades with a form of the same standardized test (one of several available commercially) which was not changed more than once or twice a decade, today it is not uncommon for states to develop their own measures, and to change these assessments annually. In addition, some states have adopted matrix sampling approaches in which different forms of the assessment are given to different students at the same grade level. In addition, states are assessing more grades and more subject areas. All of these complexities make it more difficult for states to assure accurate results on their assessments.

Both the visibility and the complexity of testing strongly suggest that states need to take a systematic approach to quality control and that they actively pursue the develop of materials and procedures to enhance the accuracy of the results they produce. The quality and accuracy of the data produced by state testing programs will be the result of the development and the implementation of these procedures and methods.

The purpose of this paper is to describe how a state can develop the procedures and the materials to help assure that the results reported are accurate and of the highest quality. They are based on the experiences of large-scale assessment programs in a number of states as the assessments were designed and implemented. Too often, the lessons shared here were "learned the hard way," but the hope is that by sharing them, others can avoid making these or similar mistakes.

The first part of the paper describes some of the ways of building accuracy in the scoring system as it is designed, developed, proofed and used. The development of the scoring system is a lengthy process, typically involving both a contractor and the agency sponsoring the work. There are several important stages, and it is important that the accuracy of the scoring system be verified at each stage.

The second part of the paper describes the documentation that is needed to assure that an accurate record of the scoring system is developed and maintained. This may be for the purpose of the initial development of the scoring system, to demonstrate that the system was designed and implemented as intended, or for future use as the scoring system is re-used in subsequent years. "Getting it in writing" is essential in assuring that an accurate scoring system is developed. Yet it is rare that such documentation is prepared accurately and completely. Critical decisions or data needed to assure an accurate scoring system may not be put in writing due both to the last-minute nature of some of these decisions, as well as the limited time with which to develop the system ("I can either develop the scoring system, or I can document it, but I can't do both") and to proof it.

This paper addresses both of these critical needs and provides guidance to staff of both the agencies sponsoring the work and the contractors hired to carry it out.

Designing and Developing the Scoring System

There are several steps that typically are used to develop the scoring system that will be used to prepare student assessment results. Each of these steps, and the quality assurance steps that need to be built into each step, is described in this chapter.

Steps in Developing the Scoring System Design

The development of a quality, accurate scoring system depends on the clear statement of purposes of the system, the accuracy of the different types of data that will be inputted, the clear determination of which types of reports the system will be designed to produce, and the manner in which the reports will be assembled and distributed. Since the typical assessment program involves students at one or more grade levels taking assessments in two or more subject areas, comprised of one or more item types, a number of decisions will need to made at the outset of the system design. Typically, these are decisions that the agency sponsoring the assessment program will make, but may need to be prompted and discussed by the contractor developing the scoring system, since there are cost and developmental implications inherent in these design decisions. Even states that are using "off-the-shelf" test products will have some custom-development of aspects of the scoring system, such as custom or adapted score reports. The extent to which the following steps apply is highly contingent on the extent of the customization of the scoring system and score reports.

Overview of the Development Process

The first step in the design process, whether after the initial award of a contract or the start of another year in a multi-year contract, will be for the sponsoring agency and the contractor to meet. This "kick-off meeting" will serve a number of purposes, as will be discussed below. Following this initial meeting, the contractor should document decisions made, the issues that need to further discussion and resolution, and issues or areas of uncertainty that have arisen subsequently.

The contractor will be responsible for beginning to design the scoring system. This design work will undoubtedly raise a number of important questions and issues, and these are often resolved informally by telephone, or more formally via memoranda, letters, and /or faxes between the contractor's project director and the contact person(s) at the sponsoring agency. These contacts help to assure that the system design follows the ideas and specifications discussed at the kick-off meeting and documented following it.

At one or more key times during the system design, the contractor staff may wish to show staff of the sponsoring agency their work in progress. These "mini-progress reports" serve to keep the state apprised of progress on the system design, head off any trouble before it is too late (or too expensive) to correct, and help the state remain comfortable with the progress of the contractor.

Once the system is completely designed, a thorough system review should take place on a pre-specified date several weeks in advance of the date by which the scoring will first be needed. At this meeting, the contractor should demonstrate the entire scoring system by having run all of the score reports, particularly those that will be sent back to school districts. The sponsoring agency and the contractor should have carefully constructed a test deck of answer documents to thoroughly check out the scoring system, and used this test deck to generate the sample reports.

Following this system demonstration, the contractor will make any changes needed in the scoring system and then await the first live data.

Once testing begins (or shortly beforehand), the state should locate a school district willing to be a "test district." This district will be offered all of the optional services (at no charge as an incentive to use them) and complete testing during the early part of the testing window provided for the assessment program. The district will then return the answer documents for scoring and careful checking. By using a test district, the state and the contractor can once again check to make sure that all scoring parameters are correctly set and being using to produce accurate results. It is not uncommon to hand-score some of the answer documents to assure that the scoring routines are producing accurate data. Once accuracy is assured, the test district can be released and other districts' data produced.

Finally, there are batch checks that should be done to assure that the scanners are accurately scanning answer documents. These may range from scanning a pre-slugged set of answer documents to edit checks built into the scoring routines as they are used to produce the data reports.

Taking care in the design, implementation, and use of the scoring system can go a long way in assuring that the scoring system accurately reports the information as desired by the sponsoring agency.

Detailed Description of the Process for Developing Accurate Scoring Systems

The process of developing an accurate scoring system involves several inter-related steps that will occur over an approximately six month schedule. The amount of work to be carried out at each step in the process will be somewhat dependent on the nature of the scoring system to be created, and whether this is the initial development of a new system, a major revision of an existing scoring system, or the updating of a scoring system to be re-used with only minor changes. The process for developing the scoring system, and assuring the accuracy of the work of both the contractor and the sponsoring agency are described below.

The Kick-Off Meeting

A key to a successful start in the design of the scoring system is the initial meeting held between the sponsoring agency and the contractor. While this initial meeting will deal with additional aspects of the assessment program beyond system design, only those aspects of this initial meeting that affect the design and implementation of the scoring system will be discussed here.

This meeting serves to lay out the sponsoring agency's reporting plans and get initial reactions about feasibility and costs from the contractor. It is also the contractor's opportunity to raise questions and issues that might have become apparent during the bidding process but which were not raised in the contractor's proposal. Finally, the meeting allows either the sponsoring agency or the contractor to suggest alternatives or additions to the planned reports.

This meeting should begin with an overview of the assessment components that the contractor will be responsible for. This should be presented by the sponsoring agency. Included in this presentation should be the sponsoring agency's ideas and designs for reports of each type to be prepared, as well as any special analyses to be prepared. Sometimes these ideas are well developed, particularly in cases where the assessment program is not changing much from the previous year. In other cases, the ideas may be quite rough, with just a general sense of what data might be reported and some uncertainty as to who will receive which reports. The contractor may wish to discuss alternatives at this point, or may wish to reserve such suggestions until they have had a chance to study the ideas of the sponsoring agency. In any case, it should be the contractor's responsibility to document the information presented in order to capture the intent of the sponsoring agency.

The contractor should then lay out plans for the development of the scoring system. This will include the timeline for the system development, the persons to be involved, the major steps in the development, and any intermediary review and/or decision points along the way. It is important for the contractor to agree upon the data to be generated, the manner in which these data are manipulated, and not only the types of reports to be generated but also the manner in which the data will be stored and made available (the storage media and the access of others to the information).

There are certain key decisions that the sponsoring agency and contractor will need to reach consensus on. These include how missing or incomplete data will be handled, whether any students are to be excluded from the reports and analyses, which aggregations of the data are to be generated (and who is to be included in these), what disaggregations are to be created and for which sub-groups, and so forth. These discussions should also include the reporting levels (student, parent, teacher, building principal, district, state, and any other), the types of reports at each level (for example, an individual student report for the student, the parent, the teacher, the permanent record), and the assembly of the reports into an overall package to return to the school or the school district.

If there is information to be provided by the sponsoring agency, or if the contractor is to provide information to the sponsoring agency who in turn generates reports or stores the information internally, these data "hand-offs" need to be discussed as well. These discussions should include the manner in which each agency will assure the other that clean information files that are the other that clean information files that are accurate will be provided. Each agency will need to incorporate steps to assure accuracy.

Following the meeting, the contractor should generate an extensive annotated memorandum of understanding about all of the information discussed at the meeting. This memorandum should document in detail each of the decisions arrived at, including the reports to be generated, the analyses to be conducted, and the developmental work on the scoring system needed. The memorandum should also spell out the issues that remain, presenting what if any decisions were made (or need to be made) and by when decisions are needed. This will serve to document the decisions made, the work that needs to be carried out, and the assignments for both agencies.

Implementing the System Design

Once the initial meeting has been held, the work on the scoring system can begin. Much of this work will be carried out by the contractor staff, working with whatever written documentation exists, the notes from the initial meeting, and any subsequent clarifications of these. Since the contractor staff will be working on this activity for some time (depending on the complexity of the scoring system, the contractor may spend eight or more weeks in system design and development work), it will be important for a couple of activities to occur.

First, the contractor will want to document in writing major decisions being made as they are made and implemented. The development of the system involves several individuals working individually, but in a coordinated manner to develop components of the scoring system. Since understanding of design decisions being made is easier as the system is being developed, it will be important for development staff to note in writing how certain key decisions were made and how these decisions were implemented. Of course, there is the natural temptation of allowing the crunch of system development to interfere with careful documentation – since it will feel like there is time only to develop the system, not to document it. However, the only record of the development of the system is that which the contractor can provide.

Second, the contractor should share the documentation it is creating with the sponsoring agency staff assigned the responsibility of monitoring the work of the contractor. By creating this documentation and reviewing it periodically with staff of the sponsoring agency, the contractor can help assure that precious time and effort are not wasted in false starts or incorrect design assumptions. Periodic sharing is particularly critical in the case of projects that are ill-defined at the outset (where the sponsoring agency did not have well-laid out design parameters at the initial meeting) or where the contractor is building the scoring system for the first time for a client.

Third, there may well be occasions during the development where the contractor will want to meet with the key staff of the sponsoring agency. For example, if the sponsoring agency had not decided on the report formats at the initial meeting, the contractor will want to rough out possible designs early in the development work, and then to review these with staff of the sponsoring agency prior to any work to implement the reports. If there are a number of these to review, or if several options have been created, it may be easier to discuss these in a face-to-face meeting, where the advantages and disadvantages of each can be presented and discussed. Other aspects of the scoring system may also be easier to review in person, where all parties involved can focus on the issues, rather than in writing, where focus on critical decisions may not occur. Of course, balancing this is that an in-person meeting takes key development staff off the developmental work and frequent meetings may interfere with the timely delivery of the scoring system. Whether or not in-person meetings are held, such designs should be approved by the sponsoring agency before being implemented by the contractor.

Documenting System Development

The written documentation developed by the contractor should be of two types. First, there are day-to-day records of design decisions. Typically, this “for-the-record” file will not be included in the final documentation generated for the program, but is available in case questions arise regarding how particular design parameters were handled, and should be referenced when the final system documentation is developed. For example, as decisions are made on the format of report forms, these decisions should be documented for the record.

Second, there are the types of documentation that will be included in the final documentation manual for the system. For example, test analyses that result in the computation of statistical scaled values for each assessment item used should be retained in a file that can be referenced and used when the system documentation is written. Each of these should be dated and then filed, so that they will be available when needed at the end of system development. The critical thing, in either case, is that while the contractor (and to a much lesser extent, the sponsoring agency) is developing the scoring system, the written record of programming decisions needs to updated constantly, so that when the system is proofed and documentation manual is written, there is a complete back-up record that provides complete detail about how the system was developed.

Steps in the Quality Control Process to Assure Accuracy

A mindset for quality, and the careful documentation of each step in the process of developing the scoring system, are important elements in developing the quality control needed to assure the computation and reporting of accurate assessment results. Since the typical scoring system can be quite complex, with myriads of look-up tables, calculations, decision rules on which students to include (and exclude) and under which conditions, as well as the wide variety of reports typically produced, on-going attention to accuracy and the maintenance of records with which to check the system are vital to assuring accuracy.

There are additional explicit steps that should be carried out. Each of these steps is somewhat dependent on the nature of scoring system and the complexity of the changes being made in it, but one generalization that can be made is that each step in the development of the scoring system should be double- and triple-checked in modular form as it is developed, then re-checked to make sure that the modules work together as designed.

A second important distinction should be made as well. While it is the contractors' contractual responsibility to deliver an accurate and effective scoring system in a timely manner, this does not in any way alleviate the sponsoring agency of the obligation of assuring that the contractor has done so. Too often, sponsoring agencies, whether due to lack of staff, staff expertise or interest, abrogate their responsibility of assuring that the contractor has performed adequately. Since it is ultimately the sponsoring agency that will be responsible for the quality and timeliness of the assessment reports, it is important for the sponsoring agency to participate at least at key points in the quality assurance process.

Third, the checks on quality should not be saved until the completion of the development work on the scoring system. The checks should be on-going, both by contractor staff and staff of the sponsoring agency. By waiting to conduct quality checks until the completion of the system development, the sponsoring agency may miss important problems, while the contractor may need to re-write sections of the scoring system (perhaps at its own cost), or needed changes may delay the availability of the scoring system (or if less critical, may not be able to be made if the scoring system has to be available when needed).

It is essential that the oversight of the sponsoring agency occur both during the time of development of the scoring system and at critical check-points along the way. In other words, the production of an accurate, timely scoring system is a dual responsibility of both the contractor and the sponsoring agency. If staff of the sponsoring agency are not available, the sponsoring agency should find other qualified third parties (e.g., local directors of testing) that can participate in the proofing/quality assurance process.

Four, while any time a scoring system is developed from scratch, or a major overhaul of it is implemented, the need for careful checking increases, there is no time when checking is not needed. Even re-mounting last year's scoring system, with no changes, can be problematic, for several reasons. Was the correct version of the system located and used? Are there any changes in operating system software that might cause problems when used with the existing scoring system? Did anyone change last year's scoring system after it was used but before it was stored for use later? Will any of the minor changes desired in the scoring system conflict in any way with the existing computer programs? These are among the questions that a system check can answer in situations where the scoring system is unchanged.

Demonstrating the Scoring System

It is recommended that an official "demonstration" date be established by which the contractor will be required to show off the scoring system to the sponsoring agency. Some sponsoring agencies may wish to hold the contractor liable for this demonstration through penalties for failure to provide the completed scoring system. These may include financial penalties that accrue on a daily basis, reductions in the overall payment, and even may include cancellation of the contract and liability for another vendor creating the scoring system on a "crash" basis.

Regardless of the sanction, if any, there should be a pre-established date by which the contractor is to hold a system demonstration for the sponsoring agency staff. This date should be established in advance, and should occur after the contractor has had a reasonable amount of time to develop the system, but with enough time in advance of the actual use of the scoring system so that the contractor can made any needed changes, or if cancellation of the contract is an option for failure to perform, another vendor has a reasonable chance of developing the scoring system before answer documents are returned for scoring.

A carefully designed "test deck" should be prepared to assure that any potential error or problem in individual student records is handled appropriately by the scoring system. The nature of the scoring system test deck will be dependent on the types of decisions expected of the scoring system. Any potential check, exclusion, or special case should be tested in a variety of ways to assure that the scoring system will handle it correctly. Each of these special conditions should, of course, be well documented in writing.

This will start with the preparation of a student file of raw information (it is typically unnecessary to actually prepare scannable documents for this, but instead, a data file specially prepared to test the various conditions is typically used). If the program relies on calculating values from live data, this test deck may consist of past year's information which is used to demonstrate the score reporting aspects of the system. The types of cases that ought to be tested, using multiple individual student records, include the following:

- entirely blank answer document;

- identification section completed only;
- incorrect test form gridded;
- first item answered only;
- last item answered only;
- imbedded omits of test items;
- last few test items left blank;
- randomly-omitted test items;
- multiple-choice items answered only;
- constructed-response items answered only;
- other conditions to be tested such as excluded students, special codes, and so forth;

The various individual student records should be "assembled" in a logical order to test the types of summary reports as well. For example, the individual student records might be separated into two or more classrooms in two or more schools, within a school district. This will allow the checking of student records at the individual student level, as well as how the individual records are aggregated to produce classroom, school, and district summaries.

A major part of the proofing the scoring system is to compare the raw files to the manner in which these and other conditions are handled by the scoring system in terms of preparing both the individual student reports and the summary reports of the results. Each special case, represented by a different individual student record, should be hand-scored and compared to the report generated by the scoring system. Then, the aggregate reports, assembled by classroom, school, and/or district should be compared to the hand-calculated summaries of the individual student records.

An essential part of this proofing is to assure that the correct answers to any selected-response exercises have been entered into the scoring system (that the sponsoring agency and/or the testing contractor have determined the correct answers or answer weights correctly), and that these values are properly stored in the scoring system and being used in processing the individual student records. In the case of matrix-sampling programs, this will be even more complex, since multiple test forms are being used and it will be essential to not only have correct answers, but to assure that the correct answer key/scoring key is being used with each test form for each student.

A "dump" of the student file should also be prepared to use in proofing the student reports. It may also be desirable to have the actual student answer documents on hand (or the scan file of the digitized images of student work), in case it is necessary to check the actual scoring of the answer document. If scoring weights are to be applied to students' raw responses, these again need to be hand-checked by manually applying the correct values to see if the scoring programs are producing accurate results. While it may be desirable to check the computer source code to see that the correct values are entered into the programs, this is not a substitute for determining that the programs can correctly look up the value(s) needed and make the correct calculations and report the appropriate values correctly.

It is critical that staff from the sponsoring agency (and the contractor) who participate in the system proofing should be familiar with the types of decisions about students that the scoring system needs to make. This proofing is both a clerical and a professional activity; since it combines both making sure that correct values were entered into scoring routines and look-up tables have the correct values in them, but also that the values entered into computer programs and look-up tables are used accurately. Hence, it is recommended that senior officials from the sponsoring agency and the contractor participate in proofing the scoring system. It helps if these individuals are the ones that can also attend to how the details of the scoring system were implemented. The team should include at least one individual who has sufficient statistical training to assure that any statistical manipulations of the information are being correctly carried out. Of course, the contractor will also have had its own quality control clerical and professional staff doing on-going system checks as the scoring system was being developed.

Documenting System Design Changes

As the system is first proofed, and re-proofed after changes are made, it is essential that all changes continue to be documented. While at no time is the scoring system development taking place on a leisurely pace, the proofing stage usually occurs just before the system is actually used. Hence, there is typically not much time to make any needed last-minute adjustments before the system is actually used, much less to document these changes. However, it is critical that these last minute changes are documented, and that any subsequent versions of the scoring system are carefully labeled and documented. In the last minute rush to prepare for actual scoring, it is certainly possible that an older, less accurate version of the scoring system may be mounted and used when actual answer documents are received. It is therefore critical to determine which version of the scoring system is the final one and to make sure that all changes to the system have been documented in writing.

Initial System Use and Final Proofing

It seems that no matter how creative the sponsoring agency and contractor staff have been in developing and using the test deck to test all conditions before the scoring system is used to produce student results, the scoring of actual student answer documents will do unanticipated things to the scoring system. In addition, last minutes changes, possible confusion over the version of the scoring system actually to be used, and so forth all point to the need to check actual live data from school districts.

These potential problems, which can best be detected from a careful examination of live information from a number of students tested in multiple classrooms in multiple schools, yet taking the time to carefully check information from the student to the district level summaries without unduly delaying the scoring of all the remaining answer documents, requires that a good-sized school district be recruited to serve as a "test district." Ideally, this should be a district with two or more schools at each level tested, should be able to take advantage of all of the optional services (which should be provided at no cost, as an incentive to use them), and which will complete testing as soon as possible during the testing window or shortly before the beginning of the testing window, so that any changes needed at this point can be made without delaying the preparation of other results. The answer documents should be over-nighted to the contractor.

Once the answer documents arrive at the contractor's site, they should be processed as all other answer documents will be processed. This includes running the materials through the check-in and initial pre-editing checks, determining missing or damaged materials, and making whatever corrections needed. It is important that the contractor not treat the test district as too much of a special case, so that all of the steps in the process that will be used with the remaining answer documents can also be tried out at this step. This will allow the contractor to identify steps in the check-in and pre-editing processes that need to be changed or bolstered. The contractor should process the answer documents, score them, and prepare a complete set of reports, from the student/parent level to district summaries. A "dump" of the student file should also be prepared to use in proofing the student reports. It may also be desirable to have the actual student answer documents on hand (or the scan file of the digitized images of student work), in case it is necessary to check the actual scoring of the answer document.

Ideally, the same individuals from both the contractor and the sponsoring agency who were involved in the proofing of the scoring system on the official check date should also be involved at this point. This also includes the individuals with sufficient statistical background and training to assure that any weightings or other statistical manipulations of the data are correctly carried out.

On-Going Error Detection

As answer documents are returned for scoring by the scoring contractor, the materials will be checked in by the contractor and any errors detected at check-in will be flagged for immediate or later correction. Errors detected at this stage could include missing or incorrectly completed header information, incorrectly assembled answer documents, and so forth.

Once the answer documents are batched and sent to processing, other errors may be detected, including incorrectly completed answer documents (either blank or incorrectly completed identification information; mis-gridded answers; damaged answer documents). These will either be corrected during the processing process or handled by the scoring system as the results are reported.

The scoring contractor will be primarily responsible for the detection and correction of errors. Any unique or unanticipated problems which need to be resolved by the sponsoring agency will need to be raised by the contractor with the sponsoring agency. Once these are discussed and resolved, the contractor will need to make the corrections agreed upon. Of course, any changes made to the scoring system will need to be proofed thoroughly. Since these will be last-minute, emergency changes, the impact of making these changes will need to be completely assessed in the short amount of time available to make them. These are proofing activities that the contractor will carry out, but the sponsoring agency staff may be involved if the changes are substantial.

Any changes to the scoring system that are made during the live scoring should be well-documented by the scoring contractor, since presumably most if not all of these changes will become permanent for future use in scoring assessment documents in subsequent years.

Post-Reporting Error Detection and Correction

Once the results are printed and distributed to local school districts, there are rare instances where errors may be detected. While this is unfortunate, these errors do need to be thoroughly reviewed immediately. If an error is suspected, an immediate halt should be placed on the distribution by the contractor of any results that have not been returned to local school districts. Second, the sponsoring agency and contractor should notify the other at once. Of course, local school districts that have already received results should be notified about the error, the steps being taken to isolate it, and when either corrected results will be available or subsequent information about report corrections will be known. It may be helpful for the sponsoring agency or contractor to contact a few local school districts known to them to see if the problem did occur in their results. This may help in isolating the extent of the problem.

A review of the scoring system should be started at once. This review should serve most immediately to do two things. First, the review should help identify what happened and under what circumstances. Is the error universal (e.g., all fourth grade mathematics individual student reports had the error?) or did it occur only in select cases (e.g., the fourth graders incorrectly answering items 27 and 28 were not scored properly). The contractor and the sponsoring agency should take steps to identify the nature and the extent of the problem. For example, is the problem limited to a particular grade or subject area? Is it limited to results produced in a particular batch? Results from a particular test form? Produced on a particular date? Both the contractor and the sponsoring agency will need to "play detective" in trying to track down the error and identifying when and how the error occurred so that the affected results can be "recalled" and the error corrected so that it does not re-occur. The review should also serve to identify whether the problem occurred with all students processed up to that point, whether it only occurred in certain batches, or randomly across students and batches.

Once the source of the error is determined, steps can then be taken to immediately correct the problem. Since these are changes to the live scoring system being made while student answer documents are being held for processing, extreme care will need to be exercised that the corrections being made don't introduce new problems into the scoring system. These corrections will be made, and proofed, on an emergency basis, so extreme care will need to be taken to carry out the changes quickly and accurately.

Corrected reports will need to be prepared and sent to school districts while additional answer documents are being processed. It is best for these corrected reports to be sent to school districts packaged in a comparable manner as the regular reports (for example, separated by school building and/or teacher), but sent with a cover memorandum explaining the extent of the error (in which reports and how much of a difference it made in those reports), the steps taken to correct it, and which reports are to be discarded and which are to be retained.

Once again, any changes to the scoring system that are made during the live scoring should be well-documented by the scoring contractor, since presumably most if not all of these changes will become permanent for future use in scoring assessment documents in subsequent years. The errors made may also point to the need for improving the proofing of the scoring system in subsequent years, whether during the scoring system development, the first system demonstration, or the verification of the first live data.

Summary

In most cases, the processing of student answer documents by the contractor's scoring system proceeds error-free and on schedule. The steps taken in the careful design and implementation of the scoring system, starting with the kick-off meeting months earlier, proceeding through the scoring system development, verified at the system demonstration meeting and further proofed when the first live data is received has assured that a high quality, accurate scoring system has been developed and is being used to accurately process all student results. The redundancy built into the system development has assured that errors have been caught and corrected when detected, and that error detection was completed before live results are produced and returned to school districts. This is a lengthy process, with many steps, and the stakes are high when errors do occur. As stated at the outset, the responsibility for assuring the development of an accurate scoring system lies jointly with the sponsoring agency and the scoring contractor. Each needs to be actively involved in designing, developing, and proofing the system.

Documenting the Scoring System

One of the most important steps in assuring the accuracy of student assessment scoring systems is the documentation of the scoring system as it is developed, used, and changed. Rarely do sponsoring agencies have in hand a complete and thoroughly checked documentation manual about their scoring system by the end of the assessment year. While the need for such a complete manual can become painfully obvious when questions arise or a scoring error occurs, when results are prepared and distributed in a manner not intended by the sponsoring agency, or when the sponsoring agency changes scoring contractors or key staff of the sponsoring agency leave, it may be too late when these events occur for the staff of either the sponsoring agency or the contractor to piece together how critical aspects of the scoring system were designed and implemented.

The development of documentation about the scoring system is an on-going activity, and often needs to be done when there is little or no time to do such work. It is not uncommon for staff involved to feel that there is either time to develop the system or to document the development, but not both. Yet, by writing down the steps used to develop the system with key data elements and look-up tables in appendices, a critical element of the development of an accurate scoring system is available for use in proofing the scoring system at each stage identified in the earlier sections of this paper, and can be used in the future in the re-development of the system in subsequent years. Hence, documentation is a valuable activity in both the short-term and the long-term.

Why is Documentation Needed?

The obvious answer to this question is that documentation is needed because human memory is not perfect and fades with time. In the short-run, different persons involved in the development of the scoring system may have different perceptions of the decisions that were made, while in the long run, key staff may not remember how the key decisions were made. The key values hand-calculated and used during programming may get lost or other similar tables found, and uncertainty arise over which table is the actual one used or how the values shown in the tables were derived. In addition, key staff involved in the decision-making process one year may not be available later in the assessment year, much less in the future. Key staff may leave the sponsoring agency or the contractor, or be assigned to other activities. It is not at all atypical for the intricacies of the scoring system to be known to only one or two persons at just the contractor, and perhaps one person at the sponsoring agency. What would happen if one of these key individual quits, is injured in an accident, or has a heart attack? How will the decisions that have been made during the development of the scoring system be known in such a way that they can be double- and triple-checked?

Almost every sponsoring agency must also periodically re-bid the project for which the scoring system has been developed. This may be after one contractor has had the contractor one or two years, or worse, after the initial contractor has had the contract for five years or more. The only thing worse than creating the scoring system in the first place is to re-create it using another vendor, assuring that the scores and score reports produced in past years will be reproduced by the new vendor so that any changes in scores from last year to this year are due to changes in student performance, not unintended changes in scoring system. This is a particular challenge since last year's contractor may not be at all pleased with losing the contract, and the winning contractor may not have completely understood the needs of the sponsoring agency. Hence, the previous contractor may not be completely forthcoming in helping the sponsoring agency move the scoring system to a new vendor, despite protestations to the contrary. The out-going contractor has little or no financial incentive to assist the new contractor in developing, much less improving the scoring system. Mix into this changes that periodically occur within the sponsoring agency staff, and it is obvious that there are a variety of ways that disaster can occur.

One final note: although sponsoring agencies may not develop complete documentation of their scoring system when it is first developed, it is never too late to do so. Although documenting the scoring system may be more challenging when it has been used for one or more years, due to changes in the system, new staff at the contractor and sponsoring agencies, and the challenge of remembering decisions made in past years, it is still valuable to put the decisions in writing, together in one place the documentation needed to re-create the system. This may mean scouring the files of either the contractor or the sponsoring agency, but writing the documentation manual is still a very worthwhile activity to carry out.

What Documentation is Needed?

This is both a very simple and a very complex question to answer. The simple answer is "everything." The more complex answer is that each and every decision to be made by machine or by people in processing answer documents, producing results, and assembling these for distribution should be written down and pulled together in one place. The documentation manual is the "one place" where all of this information resides.

The documentation manual should be organized into two major sections: the first should be a description of each test form and accompanying information and the second should be a description of the information to be found on each report produced by the scoring system and the process and decisions involved in producing this information. Each of these sections is described below.

Documenting the Tests

The first section of the documentation manual (or, perhaps in a separate program library, indexed in the documentation manual) will be complete descriptions of the tests used in the assessment program for which the scoring system will be produced. This documentation should contain the following pieces:

Test Design The documentation of the tests used should first provide a complete description of the testing design(s) used in the assessment program. For example, what types of tests are used, are all students assessed with each form or is some form of matrix sampling used, which results are to be returned and to whom, and so forth? Since the design may change in subtle (or not so subtle) ways each year, and new programs introduced or old ones phased out, it is important to understand the context of the assessment program, particularly how the assessment instruments are used. It is also important to document the tryout or piloting of new exercises, whether these exercises are a part of the regular test booklets or are printed in separate booklets. The documentation manual should provide an overview of tryouts conducted within the regular testing period and testing conducted at special times of the school year.

Test Assembly Rules The overall process for assembling the test forms should be provided, to better understand how the test design is actually operationalized. This documentation may include both a description of the design, and tables that show the numbers of each type of exercise in each section of the test(s) used.

Actual Tests or Test Forms Some programs use a common test for all students at a grade level, while others use a matrix-sampling design involving multiple forms. In either case, the documentation manual should contain a copy of each test form or assessment booklet used in the program at each grade assessed. If tryout forms are in separate booklets, these should also be made a part of the documentation manual. If the tryout exercises are imbedded in copies of the regular assessment booklet, then each copy of the regular assessment booklet with the tryout forms in them should be included in the documentation manual.

Test Administration Manual(s) Each manual used to provide tests administration directions to the persons who administer the assessment (e.g., classroom teachers), as well as who supervise the test administration at the school or district levels (e.g., , school or district test coordinators), should be included in the documentation manual. This will help to assure that any directions that affect any of the assessment items will be available in the future to understand the context of the assessments, particularly for exercise that are read to the students.

Ancillary Testing Materials There are a variety of ancillary testing materials that may be prepared that could well affect the production of reports from the assessment program. These include various student handouts, such as paper shapes, rulers, and other devices used in mathematics assessments, various header sheets (for example, at the classroom, school, or school district levels), background questionaires (e.g., to determine instructional practices), or forms used to collect information about the students assessed or not assessed. Samples of these materials should also be included in the documentation.

Answer Key and/or Scoring Rubrics For each test form used and contained in the documentation manual, the answer key and/or scoring rubrics needed to score the assessment instruments and exercises should be given. This documentation of scoring keys range from look-up tables to extensive written guides to analyzing student responses. The documentation will be needed regardless of whether the tests or assessments are selected-response, constructed-response, or other types of performance assessments.

Since the nature of the assessment and the mixture of item types will vary from one state to another, the exact nature of this documentation will also vary, but the critical element is to provide a copy of the documents needed to ascertain the correct answer or preferred answers to each test item used on each test form. In the cases where students need to produce an answer, the documentation manual should provide the reader with the rubric or guide for scoring student response.

If feasible, the samples of student work used to document each score point of the rubric or guide should be included in the documentation manual. However, it is not necessary to include within this documentation all of the samples of student work used to train scorers nor the samples used to validate the scorer training, but since these are critical to assuring that scorers are scoring the same exercises in the same way in the future, these samples of student work need to be filed somewhere in the sponsoring agency and that location(s) should be noted in the documentation manual.

File Layouts The documentation manual should contain each of the computer file layouts for each different file that the program will produce. This will include, for many programs, an individual student file layout, and a school and district summary file layout. In addition, there may be special versions of these file layouts, such as an anonymous pupil files prepared for researchers, and school and district summary files with more (or less) information on it for local districts or for researchers.

The file layout should completely define field titles or names, and describe the origins of the data contained in each field, since often very cryptic field names are used and users of the file layouts may not have access to the other documentation contained in the entire documentation manuals. It is critical, therefore, that the file layouts be completely self-contained, i.e., that they provide everything a user (particularly individuals who are not familiar with the program) need to understand and use the data files.

Documenting Each Report and Analysis

Documenting each of the reports that the scoring system produces may seem to be a relatively easy task. After all, each assessment program produces a set of reports and a copy of each can be provided in the documentation manual. This seems about as easy as providing a copy of each test booklet or test administration manual. However, much more is needed!

This section of the documentation manual will be the most complex and take the most time to compile. Each and every report produced by the assessment program, whether routinely produced or a special report, whether produced and returned to school districts or sent only to the sponsoring agency, should first be assembled in one place. This compilation of reports should include not only reports produced on pre-printed report stock or laser printed for "public consumption," but also all of the computer-printed statistical analyses that may be routinely produced or specially-produced for use by either the sponsoring agency or the contractor, whether used as input to the scoring system or to verify the technical qualities of the assessment program or its instruments. It is critical to developing complete documentation that all reports be documented.

The assessment program documentation should describe how the program's assessment exercises are to be scored, as described in the previous section. This included the answer key(s) for the selected-response exercises, and the scoring procedures for any constructed-response or other performance exercises. In this section of the documentation manual, the documentation should describe how the answer keys, the scoring rubrics and samples of student work are actually used to score students' responses. This will help to assure that the procedures to implement the scoring system are well-documented, so that such procedures can be replicated in the future to reduce the likelihood that changes in scoring procedures are not unintentionally introduced that might affect the ability to report longitudinal comparisons of results.

There are some additional subtleties that need to be covered within the context of documenting how reports are produced. These include:

Individual Student Rules There are a variety of conditions that affect if and how individual students are included in the various reports. For example, how would the following conditions be treated?

- only the identification section of an answer sheet is completed;

- a few items at the beginning of the test are completed and the rest are left blank;

- a few items at the end of the test are completed and the rest are left blank;

- a few items throughout the test are left blank, but all others ar filled in;

- one or more items are multiple-gridded;

- the selected-response section is completed, but the constructed-response section of the test is omitted;

- the selected-response section is omitted, but the constructed-response section of the test is completed;

- the student is exempted from the assessment for one of several reasons: chronic absentiism, medical reasons, LEP status, and so forth;

- the student refused to participate in the assessment, or the students'

  parents refused to permit the student to complete one or more sections of the assessment program;

- students who left the school or the state during the testing window (e.g., a migrant student who left school before the assessments were completed);

- the inclusion or exclusion of students from private schools in the public school (or private school) reports, particularly those students who are dual enrolled in both public and private schools;

- students who were ineligible for testing due to age, grade level, course credits, and so forth.

These and other conditions, and how they will be treated in developing individual student reports (for teachers or parents) and the group summaries, will need to be contained in the documentation manual.

The documentation manual will need to describe how the responses of the student to the individual test items are summarized and reported. For example, the documentation manual will need to describe any of the following different types of summaries of individual student performance by:

- content standard or by objective;

- sub-test;

- test form;

- content area;

- cross-subject area summary.

For each of these types of summaries, the documentation manual should provide complete descriptions of the manner in which these summaries are prepared. This may be narrative in nature, or may consist of one or more look-up tables. Make sure that this documentation is prepared for each test form used, including the tryout booklets used in the assessment program.

Group Summary Rules There are conditions that the testing program will need to account for when group summaries (classroom, school, school district, and state summaries) are prepared. These include:

- which students are to be included in the summaries and which excluded. The excluded students may be ones who did not attempt the assessment, who are students who are assessed but are not to be included in summaries (e.g., LEP students), who were ineligible to be tested, or students for whom an answer sheet was completed to indicate that they have been purposely excluded from the assessment;

- schools and/or school districts that might be excluded from summaries at a higher level. For example, special schools (e.g., a state school for the blind) might use the assessments "informally" and receive individual student reports, but be excluded from summary reports because such a school does not belong to a school district and are not officially a part of the school system in which the school is located;

- school sizes for which group summaries are to be suppressed, since the school size is too small and the release of a school summary may reveal the achievement of individual students;

- the manner in which summary results are to be disaggregated and for which sub-groups. These disaggregated reports may be routinely prepared and printed, or the results may be incorporated into other reports. It is not uncommon for these reports to be prepared only for schools of a minimum size and/or for sub-groups of that are of a minimum size.

Statistical Analyses and Decisions It is important that if the results are analyzed within the computer programs used to produce the reports, and decisions are made at the student level or at a summary level, such as the school, school district, or state levels, that the statistical procedures and decision rules used should be completely documented. The other commonly carried out statistical analyses are those carried out routinely at the completion of the assessment program year to summarize the technical characteristics of the assessment program instruments, either for the preparation of the program's technical report, to file for future use, or to use to improve the assessments or the assessment program.

Such documentation includes a complete written description of the step-by-step decision processes used, and may also include the actual computer code used to prepare the reports. These statistical analyses and decisions may include:

- individual student designation of performance, including whether or not the student has passed the test or at what level of proficiency the student's score places them;

- the level of performance of the school or school district;

- the change in performance of the school or the school district, including whether the school is improving or declining in performance, such as whether the school is making adequate yearly progress, whether the school is in need of special help (or exemplary), and other designations that may be applied to the school or district results by the assessment and/or accountability program(s) in the state;

- the manner in which performance across the various subject areas and/or grades assessed are summarized in an overall designation of the performance of the school or school district.

Procedures for Preparing the Documentation

Once all reports are compiled and organized in the manual, the work of pulling together the documentation can begin (although it is hoped that much of the written documentation will already exist). The easiest strategy to start this process is to begin with reports that provide individual student results and work through the system until summary reports and statistical analyses are reviewed.

For each report or analysis, it is helpful to letter each and every separate field on the report. For example, the identification information on the report should be labeled, the headers on each field should be labeled, as should the body of each table or field, and the summary at the end or side of each table or field. It is not uncommon for a one-page report to contain up to 30 or more labels.

Each label should then be defined completely (and without jargon or acronyms), and the origins and manipulations of the data needed to produce the data in each field (whether alphabetic or numerical) should be provided. In some cases (such as the name and identification number of the school), this may be quite easy to describe (e.g., "This field provides the name and identification number of the school building in which the student took the assessment. The identification information is taken from the school header sheet which the school assessment coordinator completes following assessment and places on top of the students' answer documents before submission through the district assessment coordinator to the scoring center").

In other cases, this description may be much more complex. For example, the scaled score assigned to a student may be the result of comparing each test item that the student was asked to complete (a combination of selected- and constructed-response exercises), the statistical values assigned to each assessment item or set of items, and through the use of various statistical routines and/or look-up tables, the assignment of an appropriate scaled score to the student.

The appendices to the documentation manual should provide a copy of each of the tables used in the program. For example, if there are statistical values assigned to each test item or test form, these should be added to the manual via a table or other compilation in the body of the manual (listed where the table is used for a particular report form) or in the appendix, with references in the body of the manual when used to calculate the values listed in one or more reports. These should be referred to each time they are used in each and every report that rely on these tables.

Once this section of the documentation manual is drafted, the manual should be read by someone only minimally familiar with the intricacies of the assessment program to see if this individual is able to understand the following:

- the definition of each field on all reports and analyses;

- the manner in which the data contained in each field are compiled or computed;

- the manner in which look-up tables and other analyses are used to prepare the reports.

Documents That Do Not Provide Adequate Documentation

There are a variety of documentation that sponsoring agencies or their contractors may produce, and while they are valuable, they do not meet the need for thoroughly documenting the scoring system. For example, while states will need to produce test administration or operations manuals, these will not describe the manner in which the assessment results will be reported. While materials to explain the reports which the assessment program produces to various users will be produced, these are not sufficient to document the origins of the information reported on the report forms (although the reports will be a major feature of the documentation manual).

The sponsoring agency may have also produced a test blueprint for the assessments, and while these have value in documenting the assessment design and the assessment instruments, they will not be sufficient to document the scoring system. Finally, technical reports are important to demonstrate the technical qualities of the assessment instruments to interested users, but they will not describe the process of producing the assessment program's results.

Who Develops the Manual?

The documentation manual should be a jointly-developed product of both the sponsoring agency and the current contractor. This is both a challenge and a realistic solution. The challenge inherent in this is that it is hard to jointly develop any document in which two or more individuals located in different geographical locations work on the same sections of the document, although this is necessary since both the sponsoring agency and the contractor have made the decisions and the documents that will need to be documented in the manual. In addition, each agency has developed products that will need to be included in the manual, and each has used the products to produce the reports that the assessment program has prepared.

Realistically, however, either the sponsoring agency or the contractor should write the initial draft of the documentation manual and should write as much of the manual as possible, leaving sections to be added by the other party where the other party is the most logical source for that information. The other party should then review the originating party's work, noting changes from the other party's perspective, and adding missing sections. The originating party should then re-review the manual and note differences in decisions or other issues to be resolved. This can then be resolved and the final version of the documentation manual prepared.

Updating the System Documentation

The documentation manual should be prepared as a final product of the assessment program on an annual basis. While preparing and updating a documentation manual is not exactly one of the most exciting activities in an assessment program, and may occur at a very busy time of the assessment program cycle, it is important to carry out in the first year of the assessment program, and to update the manual annually at the end of each assessment program year. Of course, unless the assessment program changes dramatically (which does occur periodically), the work to be done annually will be consist more of updates to the information (new versions of the tests, new answer keys, new look-up tables, revisions to the report forms, and new or revised decision-making rules used to make decisions about students or schools) rather than a completely new manual.

However, it is important to resist the temptation to not update the manual annually when the changes are not too dramatic, since waiting for several years to bring the manual up-to-date may mean that the manual is out-of-date when the program contractor changes or serious issues arise about the manner in which assessment results are prepared – the very reasons why complete documentation can be invaluable.

Who Updates the Documentation Manual?

The sponsoring agency can assure that the documentation manual is updated annually by making the annual activity a contingency for final payment to the contractor for work on the assessment program. This will provide an incentive to both the sponsoring agency and the contractor to produce and update the manual before the next assessment program year begins (and before the decisions made in the previous assessment year are forgotten or subsequently changed).

This may not work, however, in cases where the contract is a multi-year one with on-going payments to the contractor not related to the completion of work. In these cases, it will be important to set a date by which the documentation will need to be completed.

Storing the Documentation Manual

Once the documentation manual is drafted, reviewed, and completed, major parts of it will be historical in nature and will probably not be re-used. For example, different tests may be used, with different answer keys and scoring rubrics, so that these portions will definitely need to be updated each year. This means that a safe (and probably secure) location will need to be identified for storing the documentation manual produced each year. While this is a working document, used on an on-going basis to assure continuity of the scoring system from year to year, it is also an archive, serving to provide a historical record of what was administered and how this was scored and reported each assessment year.

At first, the advantage of this historical record may not be apparent. However, if the program remains basically the same over a period of years, questions may arise about how particular situations or assessments were handled in the past, at a time when staff of the sponsoring agency or the contractor may not be the same. This historical record may be the only link with the past, the instruments that were then used, and the manner in which these were administered, scored, and/or reported.

In all likelihood, this means making multiple copies of the documentation manual. At least three copies are needed: one for the file (to be stored in the safe, secure location previously identified), and one each for the key staff person from the sponsoring agency and the contractor. It is important, however, to assure that all copies are completely identical, right down to the samples of testing materials and report forms. It is also critical that each page of the documentation manual be dated, so that minor text changes are not missed in a casual examination of the document, and so that copies of pages can be readily identified as to the assessment year to which they pertain and the date on which the page was prepared or modified.

How Will the Documentation Manual be Used?

There are several ways in which the assessment program documentation manual can and will be used. These include the following:

- responding to questions regarding the assessment program during the assessment year itself;

- responding to questions regarding the assessment program in subsequent assessment years;

- to assure that the assessment program procedures and processing rules do not unintentionally change from year to year;

- when the assessment contractor changes, to assure that the assessment program procedures and processing rules do not unintentionally change from the previous year to the current year.

- serving as a historical record of the program design, the assessment materials, the scoring procedures, and the reports produced.

Changing Contractors

When the agency sponsoring the assessment program changes contractors, the value of the documentation manual will become readily apparent, particularly if the sponsoring agency has held the out-going contractor to the requirement of updating the documentation manual one last time before final payment is made on the scoring contract. Assuming that this has taken place, the sponsoring agency and the incoming contractor can use the "kick-off" meeting to review the existing assessment program and any changes the sponsoring agency wishes to make in it, using the documentation manual as a source of information about the existing assessment program and the reports and analyses that it produces.

The advantage of the documentation manual is that it can serve as the focus of the initial meeting between sponsoring agency and the new contractor, since it provides detailed information about the past program and the manner in which the program has produced its reports. Should the change in contractors coincide with major changes in the assessment program, the documentation manual will be of less value, but still can provide the new contractor with a perspective on what was done in the past and how it is changing in the upcoming year.

By the end of the assessment year, the new contractor will have had to not only update the scoring system but also have prepared the revised documentation manual. Throughout the assessment year, as the new contractor's work is reviewed and proofed, reference should be made back to the the documentation manual, to help assure that no unintentional changes are made to the scoring system, and that on-going reports are prepared in a comparable fashion by the new contractor. It will also be important to assure that the updated documentation manual does not introduce unintended changes to the documentation nor to the scoring system.

Should the scoring contractor be terminated due to issues that arise during the assessment year, and a new contractor be hired during that assessment, it is still preferable to have the out-going contractor update the documentation manual if at all possible before final payment is made to the out-going contractor. Since continuity is critical, the more that can be done to provide a written record of the changes made in the scoring system since the last assessment program cycle ended, the better. In these cases, particularly when contract termination was related to issues arising during scoring and reporting, the written documentation should be thoroughly proofed for accuracy and completeness.

Summary

The documentation manual is the key to assuring that a quality scoring system is set up and implemented initially and that the system is maintained each year. By updating the documentation manual annually, as the final activity of the assessment year, the sponsoring agency can assure that the decisions and reports used in the assessment program are thoroughly described and documented. This documentation will be important not only to assure the continuity of the program from year-to-year, it will be essential when the assessment contractor changes. By making the updating of the documentation manual a closing activity before final payment is made, the sponsoring agency can help to assure that the documentation manual is updated annually, particularly when the contractor changes.

Summary

The first part of the paper described some of the ways to build accurate scoring systems as the scoring system is designed, developed, proofed and used. The development of the scoring system is a lengthy process, involving both the contractor and the sponsoring agency. There are several important stages, and it is important that the accuracy of the scoring system be verified at each stage.

The second part of the paper described the documentation that is needed to assure that an accurate record of the scoring system is developed and updated. This may be for the purpose of the initial development of the scoring system, to demonstrate that the system was designed and implemented as intended, and for future use as the scoring system is re-used in subsequent years. "Getting in writing" is essential in assuring that an accurate scoring system is developed. Critical decisions or data needed to assure an accurate scoring system may not be put in writing due both to the last-minute nature of some of these, as well as the limited time with which to develop the system.

The purpose of this paper was to describe how a state can develop the procedures and the materials to help assure that the results reported are accurate and of the highest quality. This paper addressed both of these critical needs and provides guidance to staff of both the agencies sponsoring the work and the contractors hired to carry it out.

The visibility and the complexity of testing strongly suggest that states take a systematic approach to quality control and that they actively pursue the develop of materials and procedures to enhance the accuracy of the results they produce. The quality of data produced by the state testing program will be the result of the development and the implementation of these procedures and methods.