Development of a Contract Quality Assurance Program Within the Department of Transportation

CHERYL LYNN

To assure the quality of construction products and processes, the Virginia Department of Transportation has established three levels of construction control. First, contractors themselves provide oversight and quality control as set out in the Department's Road and Bridge Specifications and in their contract stipulations. Second, the Department's construction inspectors provide quality acceptance by determining whether contractors are adhering to specifications and either accepting or rejecting the work in progress. Finally, the Department's quality assurance program examines the methods for inspecting projects and determines where new procedures are necessary to keep the construction process under departmental control. In 1987, the quality assurance effort for highway construction in Virginia was redesigned and retitled the Contract Quality Assurance Program, and steps were taken to put the program on a sound statistical footing and to improve its reliability among field personnel. Sample size requirements were calculated, and a stratified random sampling plan was instituted. A list of inspectable items was developed and prioritized to assist inspectors in managing their time and to ensure agreement on what field inspection entailed. New reporting procedures were developed to change the focus of the program from "inspecting inspectors" to evaluating the inspection process, thus removing the punitive aspects of the previous program.

The Virginia Department of Transportation (the Department) historically has been both active and diligent in its efforts to ensure that high-quality materials and efficient processes are used in the construction of its facilities. To ensure the quality of materials and workmanship, the Department has employed for many years a large complement of field inspection forces. At its inception, the Department's field inspection program provided almost all the necessary engineering services to contractors and, although it no longer provides them, it still provides comprehensive oversight of all construction and most maintenance activities. In 1965, as part of an evolutionary change in most roles for both the Department and the FHWA, the Department began its Inspection-In-Depth program (IID) in an effort to assume some of the responsibility for the construction inspection oversight previously handled by federal employees. In the most general sense, the purpose of both basic inspection and IID was to ensure the accuracy, adequacy, and effectiveness of procedures, methods, controls, and operations used by the contractor and the Department related to the construction of highway facilities (Procedures Memorandum, Management Services Division, January 1, 1977).

In 1986, a study of the statistical adequacy of the 20-year-old IID program was begun by the Department's Management Services Division, which is responsible for quality assurance. As a result of the initial study, it was decided that the original conception of the program no longer met all of the Department's needs and that a new Contract Quality Assurance Program (CQAP) should be designed and implemented. The IID's goal of providing careful and efficient oversight of the construction process was applied to the design of CQAP.

Virginia Transportation Research Council, University of Virginia, Observatory Hill, Charlottesville, Va. 22903.

BACKGROUND

Three generally accepted components, or responsibilities, are necessary for meeting the technical and legal requirements in testing and inspection: (a) quality control, (b) quality acceptance, and (c) quality assurance (see Figure 1).

Quality control is the responsibility of the contractor, who must institute procedures to ensure that the results of a project meet certain specified standards (H. Newlon, personal communication, October 1986). Quality acceptance is the responsibility of the buyer (the Department). Quality acceptance addresses both testing (attributes) and inspection (workmanship). The adequacy of a contractor's quality control can be determined by statistical sampling and analyses, particularly for those elements that involve attributes. This level of quality acceptance is, in effect, the second line of defense (assuming that a contractor has adequate quality control). Quality acceptance through inspection of workmanship is, in effect, the first line of defense for workmanship.

Quality assurance generally refers to procedures for validating the effectiveness of the combined quality control-quality acceptance system. This is the component to which the IID program is directed. In effect, this program has become the second line of defense for workmanship, just as testing is the second line of defense in the case of attributes.

These three components of quality monitoring have always been much in evidence on departmental construction sites. Until the formal development of quality control, acceptance, and assurance systems in the 1960s, the Department's project personnel often performed all of the quality acceptance functions and significant portions of the quality control functions. On federal-aid projects, quality assurance functions were provided by the FHWA through random sampling (*1*). Later, this function was performed by the Department for the FHWA. The IID program was established following a reduction in on-site inspections and tests by the FHWA.

Responsibility for the various aspects of quality acceptance

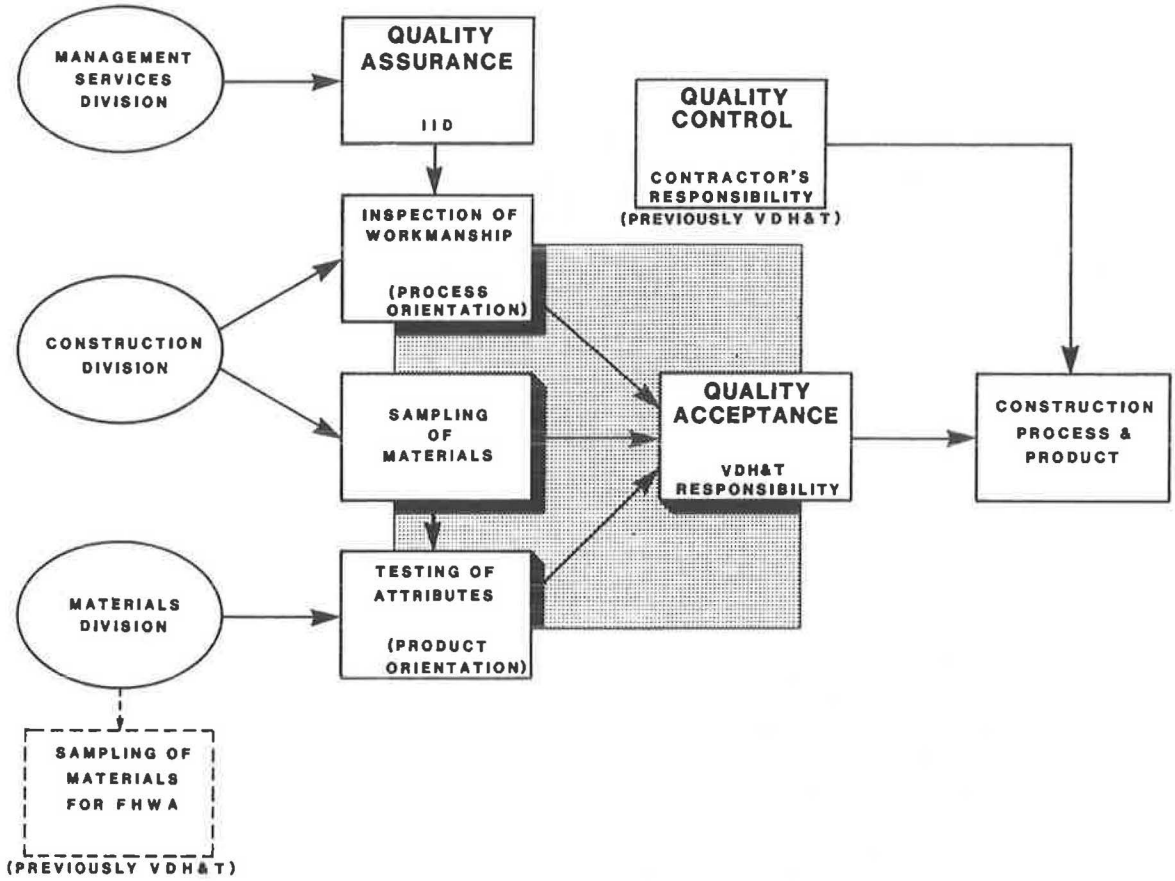

FIGURE 1 Department's quality control, acceptance and assurance system.

and quality assurance is vested in various divisions in the Department. The inspection of workmanship, mainly a process-oriented activity, is under the control of the Construction Division. The construction inspectors also provide samples of materials to the Materials Division, which has the responsibility for testing them to ensure that the materials meet specifications. The IID program, an additional control for both inspection and quality acceptance, operated from 1965 to 1986 under the auspices of Management Services and its predecessor divisions.

For several years before its intensive study of quality assurance, the Management Services Division had been attempting to improve IID site visits and site reports to make them more useful to the Department. Studies of the IID program had identified a number of statistical and procedural problems that had to be rectified before the program could adequately assess the quality of current practices. In an effort to evaluate the statistical accuracy of IID and to explore new uses for the data, both Management Services personnel (Statistical Analysis of the 1985 IID Program. Internal memorandum, Virginia Department of Transportation, 1986) and Turshen and Miller from Virginia Commonwealth University (*2*) conducted studies using the previous years' IID results. They concluded that there were no significant correlations between such variables as number of inspectors per project or dollar value of the project, which had been thought to be related to project complexity, and the number of deficiencies noted by IID inspectors. They also concluded that IID sample sizes were too small to allow for statistical analyses and that the lack of random sampling reduced the extent to which IID results could be generalized. As a result of these studies, in August 1986, Management Services requested the assistance of the Virginia Transportation Research Council (the Council), the Department's research arm, in evaluating and revamping the program.

Initially, the Management Services' request dealt exclusively with IID (quality assurance). As originally intended, the joint effort by Management Services and the Council constituted an attempt to improve the existing quality assurance program by using the same techniques that Management Services had applied successfully elsewhere in the Department. However, in examining the IID program, Council personnel became increasingly convinced that the success of IID was dependent on the success of the Construction Division's inspection program. These two programs were so interrelated that the former could not be adequately evaluated without aspects of the latter being considered. For this reason, it was proposed that the evaluation examine both programs.

The request from Management Services came at an auspicious time. Both IID and the construction inspection program had shown great potential for providing information crucial to the proper management of the construction process. The generation of this information and the monitoring and oversight of the quality acceptance program it represents were especially important in view of the significant expansion that occurred in the state's construction program in 1987 and 1988. Clearly, for the Department (and the state) to be sure that it was getting what it paid for, these new projects had to be carefully monitored to ensure adherence to contract provisions and state specifications. Also, since the Department was committed to getting many of these projects under way very

quickly, the success of each inspection activity became even more critical.

The evaluation was timely for several other reasons. Because the new construction effort was the most significant undertaken by the Department in many years, the size of the inspection force had to be greatly increased, and any needed changes in the current inspection program that might be revealed by the evaluation could be initiated before new employees became entrenched in practices that might not be critical to ensuring quality. An analysis of what was to be inspected, how it was to be inspected, and how the data obtained were to be recorded was appropriate. New procedures for evaluating the importance of various construction elements could be developed without as much opposition from the inspection establishment since most of the Department's construction practices were in the process of change and adaptation to the new work load. Thus, IID quality assurance practices and procedures could be evaluated to ensure that they concentrated on items most critical to good inspection.

RATIONALE BEHIND CONTRACT QUALITY ASSURANCE

The original IID program had two separate functions. First, IID was designed to provide quality assurance; that is, inspections were carried out to ensure that the Department's construction inspection process provided the proper oversight (as defined by the Department) of all projects. There seemed to be some disagreement as to whether the IID program was designed to evaluate the performance of the inspection system or the performance of individual inspectors. Because of this disagreement, IID was often seen as adversarial in its relation to inspectors in the field. In addition, IID was originally designed by the FHWA to provide "a thorough on-site review and evaluation of a specific contract item or combination of items which constitute a significant step in the construction process" (*1*). Thus, IID originally provided a more intensive level of inspection since it monitored both the construction and inspection programs.

In terms of intent, there are several similarities between CQAP and the old IID program. The CQAP provides intensive inspection and may assist the on-site inspector in cases in which the contractor is hesitant to correct a deficiency. CQAP also oversees the inspection process as implemented by construction inspectors just as IID did.

However, there are several important differences between IID and CQAP. First, since IID was sometimes misused to pass judgment on individual inspectors or individual field managers, field personnel did not welcome such scrutiny and rarely made use of the results. Emphasis within CQAP is deliberately drawn away from the individual inspector and toward control of the construction process. The purpose of the program is first to determine whether individual components of the construction process meet departmental specifications and, if they do not, whether corrective action is being taken. By highlighting control of the process, CQAP can be used to improve efficiency and evaluate construction activities, specifications, and training. The only instance in which CQAP may be used to judge individuals in the future is if it is eventually used in determining whether individual contractors prequalify to bid on Department construction projects.

A second difference between IID and CQAP is that CQAP is statistically based. Statistically, the number of IID inspections performed each year was relatively small. Thus, little could be said about the efficiency and accuracy of inspections. Also, to be able to allow statements about the inspection program as a whole, the sites inspected would have to be randomly selected (or at least randomly chosen within previously established strata). The method by which projects were chosen in the original IID program was biased toward the selection of projects with substandard performance. This criterion was chosen to ensure that projects with problems received attention since budgeting restrictions limited the number of possible IID inspections. Sample sizes for CQAP were carefully calculated to ensure statistical validity, and sites were selected randomly. Finally, under IID, there seemed to be little explicit agreement between field inspection and IID personnel as to what an inspector's job actually entailed and what constituted good performance on the part of an inspector. The inspection process, then, was not defined comprehensively, and decisions concerning the concentration of effort were left to an inspector's discretion. Since IID might not be measuring what inspectors actually did, IID visits were viewed suspiciously by inspectors. The first step in developing the CQAP involved developing a comprehensive list of inspectable items and activities so that both inspectors and CQAP reviewers would agree on how the construction process should be monitored.

PURPOSE AND SCOPE

The primary purpose of this project was to examine the rationale behind the old IID program and to put the new CQAP on a sound statistical basis. Fulfilling this objective also involved examining the policies and procedures used in administering the program.

To conduct an evaluation of IID, an examination of the current field inspection program was necessary. Thus, a secondary purpose of the study was to evaluate the content of the current inspection program, inasmuch as its content would affect the success of a restructured CQAP program.

METHOD AND RESULTS

Clearly, this project was an ambitious one, requiring the cooperation of many divisions and individuals. In this section, the steps taken to develop and implement the new CQAP are described along with the results of each step. The study was conducted in ten steps:

1. Notification of Department personnel that both an evaluation of the inspection program and the development of a new quality assurance program were being undertaken, and solicitation of input.
2. Development of a comprehensive list of inspectable highway construction items.
3. Prioritization and weighting of inspectable items.
4. Validation of the prioritized list of inspectable items.

5. Development of a statistical sampling plan for the program.
6. Development of computer software to implement checklists.
7. Development of a checklist scoring system and report formats.
8. The hiring and training of CQAP reviewers.
9. Development of measurable characteristics of priority inspection items.
10. Development of new policies and procedures for inspection and CQAP.

Notification of Department Personnel That an Evaluation of Inspection and IID was Being Undertaken and Solicitation of Input

As a first step in the process, the Management Services Division developed a procedures manual for the new CQAP project (*3*). Copies were circulated among all levels of management and among representatives from all field positions dealing with inspection and quality assurance. Comments were received from reviewers, and the procedures manual was amended accordingly. At the same time, the nine district engineers (the highest level of field managers) and their assistants were briefed by Management Services personnel.

Development of a Comprehensive List of Inspectable Highway Construction Items

It is clear that as the complexity and number of the Department's construction projects have increased, the role of an inspector has also become more complex. Up to this point, there has been no real effort to define comprehensively in operational terms an inspector's job.

The major tool of inspectors, in addition to the Department's Construction Manual (*4*), is the Department's *Road and Bridge Specifications* (*5*). This reference book supplies information on all types of construction and is extremely complex, covering many topics that describe the precise engineering standards used in the construction program. The *Construction Phase Inspection Handbook* (*6*), on the other hand, gives a broad outline of activities that inspectors may follow in doing their jobs, but it covers neither what items are to be inspected during those activities nor what constitutes a satisfactory inspection.

After examination, it appeared that there was a clear need for a document falling between the *Road and Bridge Specifications* and the *Construction Phase Inspection Handbook* in terms of comprehensiveness and specificity. This new document (or checklist) was envisioned as containing those inspectable items that applied to the project at hand, expressed in measurable terms, so that inspectors could clearly determine whether criteria had been met. Items to be included for a particular project could be standardized by project type and then tailored to meet the needs and characteristics of individual projects. Some indication could also be given in this document as to which of the applicable items was more or less critical to construction quality. A listing of these critical items for each project would give inspectors guidance in managing their time and in concentrating on the most crucial items.

With this intention, an attempt was made to create a list of inspectable items. Initially, personnel from Management Services who had extensive experience with both the inspection process and the IID program converted the *Specifications* into detailed inspection items, incorporating the standards to be met in each item as listed in the *Specifications*. Next, since there were obviously too many items with which any individual inspector would be familiar, items were grouped by standard areas and by physical items to be inspected. When this process had been completed, a checklist of approximately 1,800 items covering all aspects of the construction process had been created. These 1,800 items fell into approximately 75 categories (a portion of a checklist covering one category, Excavation and Embankments, appears in Table 1). Although 1,800 items is a large number of items, it was felt that in a checklist this number was manageable since not all projects would involve all of the items and not all inspectable activities would occur simultaneously.

Prioritization and Weighting of Inspectable Items

Inspectable items do not have equal potential to reduce or improve construction quality. Clearly, inspectors make determinations concerning more important and less important inspectable items on a daily basis as part of their job. It was felt that a method of making these decisions based on a consensus of inspectors and managers was needed. The Council frequently uses a modified Delphi technique when diverse groups need to achieve a consensus, especially with regard to prioritization. It includes aspects of the Delphi procedure and the nominal group technique (NGT). It is felt that the combination of these two methods results in an efficient, objective, and fair technique for prioritization. A hallmark of the Delphi technique is that it is applied exclusively by mail by the administration of several rounds of questionnaires. In each new round, the results of the prior round are provided along with an indication of the overall group assessment. Respondents remain anonymous and are free to revise their previous judgments. The purpose of the feedback is to induce a consensus. The method succeeds if there is convergence on one or several recommendations.

A major defect of the Delphi procedure, however, is that because the participants never meet face to face, there is no opportunity for the direct interchange of ideas or clarification of the issues by fellow participants. The interaction and explanatory materials are controlled by the group coordinators, and this magnifies their influence on the group outcome. Most face-to-face meetings use the usual committee approach, which has been shown to be a largely ineffective method of achieving a consensus. This approach often inhibits discussion because dominant individuals may exert a disproportionate influence on the deliberations, and members may contribute according to their self-perceived status. Additionally, members often make covert judgments but are reluctant to express them as overt criticisms because of the social pressure to conform. Finally, maintaining the group relationship requires a good deal of time and effort, which reduces the group's ability to deal with substantive problems and consider alternatives thoroughly.

NGT provides for structured interaction among participants without allowing the adversarial interchanges often seen in

TABLE 1 TYPICAL CHECKLIST OF INSPECTION ITEMS: SECTION 303–EXCAVATION AND EMBANKMENTS

_____	1.	Are embankments being placed uniformly and with the specified lift thickness? (PS = 1)*
_____	2.	Is the embankment being compacted uniformly and rolled to the outside of the fill? (PS = 1)
_____	3.	Is a density test conducted on embankment material with a fill length less than 500 feet at least every fourth 6-inch lift? (PS = 2)
_____	4.	On fills from 500 to 2,000 feet in length, is a density test conducted on at least every 10,000 cubic yards of embankment material and on every 6-inch lift within 5 feet of subgrade? (PS = 2)
_____	5.	Do the grades appear to be within the tolerances specified in Section 303.16 of the Specifications? (PS = 3)
_____	6.	Has the contractor obtained prior approval from the engineer before opening up a waste area or borrow pit? (PS = 3)
_____	7.	Has the topsoil stripping been confined to the area over which excavation is to be actively prosecuted within 15 days following the stripping operation? (PS = 3)
_____	8.	Do the records show the amount of unsuitable material removed and the person that authorized the removal? (PS = 3)
_____	9.	If the contractor used any borrowed material prior to the use of all regular excavation, did he obtain the permission of the engineer? (PS = 5)
_____	10.	Have all loose rocks larger than 4 inches in diameter been removed from the cut slopes? (PS = 5)
_____	11.	Were slopes widened or flattened to improve the safety of the project before any material was wasted? (PS = 5)
_____	12.	Was authorization given in writing to allow the contractor to waste or sell any surplus material? (PS = 5)

*PS = priority score. Highest priority items receive a 1 and the lowest receive a 5.

group discussion. Since anonymous balloting is used to rank projects, the false consensus that is often seen with voice voting is avoided, and the perception of fairness among participants is maintained.

The modified Delphi technique used to prioritize inspectable items involved several steps:

1. Final preparation of lists of inspectable items. Since it was felt that each Delphi panel would be able to consider only about 600 items at one time, each list of inspectable items developed by Management Services' personnel was composed of three subjects: (a) concrete, asphalt, and soils; (b) environment; and (c) general construction (including contract management). Each panel's checklist included up to 20 sublists covering discrete content areas.

2. Selection of the Delphi panel. Panelists were selected based on their expertise in the particular topic covered by the panel and on their level of interest and their willingness to express opinions concerning prioritization. Each panel included members from several divisions and geographic areas; thus, all aspects of the items could be discussed. Each of the three panels had 8 to 15 members, including persons who had been selected to later become the six CQAP reviewers. Thus, the Delphi process was used to help reviewers become familiar with the inspector checklists.

3. Mail-out round. The first round of rankings, in which all 600 of each panel's items were initially included, was conducted by mail so that participants could consider each item carefully. Panelists were asked to rate each item on a 1-to-5 scale, in which 1 indicated that the item was extremely important in assuring a quality construction product and 5 indicated that the item was extremely unimportant.

The phrase "assuring a quality construction product" was defined as including construction practices that affect the timely continuance of work and those directly affecting the quality of materials and the end products. From this initial round of rankings, about one-fifth of the items were given a 5 and were excluded from further consideration (unless panelists requested that an item be reincluded). In this way, each round of rankings created the lowest priority level, leaving all remaining items open for consideration in the remaining upper priority levels.

It was also decided that items in each sublist would be ranked independently rather than compared with other items

in the sublist. There were about 1,800 items under consideration. With ranking within sublists, about 375 would have been ranked as top priority, a number that could be beyond the capability of an inspector to inspect.

4. In-person panel session. After the first mail-out round of rankings, panelists met face to face to consider their lists of inspection items. At the beginning, the purpose of these meetings was described, and panelists were allowed to ask questions. The results of the mail-out round were presented, and the panelists were asked if they wished to salvage items that had previously received the lowest priority rating (and thus would not be considered in future rankings). Panelists were then asked to rerank the remaining items by selecting their top (highest-priority) and bottom (lowest-priority) 40 items. It was hoped that three rounds of ranking could be completed during two 3-hour sessions. However, because of the difficulty in selecting the rankings and scoring the results, the best any group did was two rounds (the other two groups completed one round each). Thus, the remaining rounds were completed by mail.

5. Additional mail-out rounds. Based on the discussion at the in-person panel sessions, several groups requested the opportunity to combine and re-edit items. Thus, the first mailing after the panel session solicited this input. In subsequent mailings, members were asked to rerank remaining items (for instance, in the third round of rankings, panelists were asked to prioritize their top and bottom 30 items).

After all ranking had been completed, the lists were reassembled and reordered based on their priority ratings. Panelists were then asked whether they agreed or disagreed on the priority of each item once they had been able to examine the checklist in its entirety. If someone disagreed with an item's ranking, all members were repolled, and if a majority concurred, the priority was changed. This occurred in very few cases.

Validation of the Prioritized List of Inspectable Items

After prioritization by field employees and midlevel management, validation of the rankings by upper-level management was needed. It was possible that field management had stressed activities that top management would not. For instance, in initial rankings, some field personnel gave a low score with regard to ensuring quality construction to items dealing with mandatory contracting with disadvantaged business enterprises. If this trend had continued throughout the rankings, or if all panelists had given these items a low score, these priority rankings may not have had management's approval.

For this reason, the list of prioritized items was sent to the top management in the construction and engineering areas. After lengthy consideration, these top managers accepted the priorities without change.

Development of a Statistical Sampling Plan for the Program

To correct one of the major flaws in the previous IID program, the new quality assurance plan had to be statistically valid. This meant that a statistically adequate number of samples had to be drawn in a random fashion, representing the relevant characteristics of the construction program.

The first step in developing a sampling plan was to determine what questions needed to be answered through the analysis. This step involved deciding whether the Department was interested in simply estimating the true state of affairs or whether they meant to compare conditions among various groups or years. Management also had to decide on the level of confidence to be used and the degree of accuracy or resolution required. Additionally, if comparisons were to be made, a decision had to be made whether the direction of the difference was important (e.g., whether one group is higher or lower than another rather than just different). There are different (but related) formulas for each hypothesis or analytical situation, and each question had to be associated with a different sample size or sampling technique. In general, the more complicated the question one wishes to answer, the more sophisticated the sampling plan must be (and often, the larger the sample).

A limited set of questions was proposed to be answered for the program's first year, with the intention that the number be increased in subsequent years. Initially, the following were the three questions to be answered:

1. What is the average number of deficiencies per construction project in the Commonwealth of Virginia?
2. What is the average number of deficiencies in each district?
3. Will the number of deficiencies measured next year be significantly different from those this year? (This question, of course, could not actually be answered until the next year, but the current year's sample had to be drawn so that it would be possible to answer the question at that time.)

The factors that go into determining sample size are variability (a factor that is inherent in the data and thus is out of the control of the experimenter), accuracy, and confidence (factors that are chosen by the analyst and thus are within control). These factors are represented in the following formula (*4*):

$$n = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 (SD^2)}{(M1 - M2)^2}, \text{ corrected } n = \frac{Nn}{(N + n)}$$

where

α, β = the confidence levels chosen;
$Z_{1-\alpha}, Z_{1-\beta}$ = the normal curve values corresponding to the chosen levels of confidence;
SD^2 = the variability estimate used;
$(SD^2_1 + SD^2_2)$ = two samples to be compared;
$M1$ = the true mean;
$M2$ = the estimated mean from the sample;
$M1 - M2$ = the minimum meaningful difference that one can tolerate between the sample mean and the true mean, or the minimum difference one can detect between two samples, for example, 1987 and 1986 (this factor represents accuracy or resolution); and
$\frac{Nn}{N + n}$ = a finite population size correction factor, where n is the uncorrected sample size and N is the population size.

For each question the program was designed to answer, a sample size was calculated and distributed proportionally across

the nine construction districts and nine project types. Then, to ensure that the study was statistically capable of answering all the questions, the largest sample size for each district and project type was selected (see Table 2).

As these sample sizes were applied to the program, a number of problems were noted. First, sample sizes were proportionally very large. This was due to the very high variance estimates taken from the 1985 and 1986 IID results. (Actual variances within the new program have been noted to be much lower, thus reducing future sample size requirements.) Second, since the inspectors were not actually taking data until the fourth quarter of 1987, the sample based on the second, third, and fourth quarters (i.e., the 1987 construction season) was not wholly applicable. In addition, by the fourth quarter, many of the projects listed as ongoing were actually completed, reducing the population size. Also, since projects were awarded based on type, several project numbers could be assigned to each job. For instance, one job might include widening, resurfacing, and repairing a bridge. Selecting project numbers necessitated that an inspector visit a job site more than once with no assurance that the project activity he or she wished to inspect was actually ongoing.

The following changes were recommended for the 1988 CQAP sample:

- Jobs rather than project codes should be selected as the unit to be sampled. Thus, the inspector could visit a project and inspect all ongoing activity at once.
- The samples should be pulled on a quarterly basis, rather than annually, to ensure that all projects selected for a particular quarter were ongoing.
- Sample sizes could be reduced based on more accurate variance estimates.

Steps 6–8 of the Study

Steps 6–8 of the study (development of computer software to implement checklists; development of the checklist scoring system and report formats; and hiring and training CQAP reviewers) were initiated as part of the actual implementation of the program. Software was developed to allow all CQAP reviewers to use lap-top computers to evaluate projects and print results after reviews of local construction managers and personnel. Data collected on reviews could then be sent to Richmond in computer-readable form for easy entry into the statewide analysis system.

The scoring system originally selected for use with the checklists required each reviewer to score each item on a scale

TABLE 2 PROJECTS TO BE SAMPLED IN 1987 BY DISTRICT AND CONSTRUCTION TYPE

	DISTRICT								
TYPE	BRIS	SALEM	LYNCH	RICH	SUFF	FRED	CULP	STAUN	NVa
Previously Ongoing	31	18	9	19	30	8	13	7	18
New Construction	2	4	3	10	7	1	1	3	4
Recon-struction	58	38	39	27	10	30	22	40	17
Widening/ Resurfacing	1	4	1	2	3	1	1	0	3
Resurfacing Only	2	0	7	2	1	0	0	0	0
Bridge-New/Major	10	10	7	7	8	3	3	7	4
Bridge-Rehab/Repair	3	1	5	2	1	1	2	1	1
Safety	7	8	5	13	12	9	6	9	11
Misc.	4	1	0	2	2	0	1	4	2
TOTAL	118	84	76	84	74	53	49	71	60

TOTAL = 669

from 1 to 5 where 5 was excellent, 4 was above average, 3 was average, 2 was below average, and 1 was unacceptable. There was, however, some misinterpretation of the highest scores, which led to consistent disagreement between reviewers and low interrater reliability. (Essentially, reviewers had difficulty relating the excellence of the work to meeting contract requirements. Under previous definitions, meeting specifications constituted excellent performance.) After two quarters of data collection under this system, the scoring system was modified and compressed to a four-point scale where 4 meant "exceeds contract requirements," 3 meant "meets contract requirements," 2 meant "below contract requirements," and 1 meant "unacceptable." Interestingly, exceeding contract requirements is not always advisable. For instance, placing reflectorized barrels on 12-ft centers in a work zone exceeds contract requirements; however, it also results in higher costs for the Department.

Once procedures were in place, CQAP reviewers were selected from among the Department's most experienced inspectors. In addition to their training period, these reviewers also participated in the modified Delphi process, allowing them to become more familiar with the checklists and to influence priorities based on their extensive experience.

Steps 9 and 10 of the Study

The implementation of steps 9 and 10 (development of measurable characteristics of priority inspection items and development of new policies and procedures for inspection and CQAP) is just beginning. Management Services has convened an advisory group of experts in construction practices and management to provide guidance for the program. This group will also assist in the refinement of checklist items and development of operational standards for each inspectable item. This guidance should also help CQAP reviewers achieve consistent and reliable ratings from all reviewers and all projects.

DISCUSSION OF RESULTS

Although this system is still in its infancy, its uses are already becoming clearer. As an objective and statistically based program, its results can be used to pinpoint areas in which the Department's construction process is "out of control" and assist in evaluating solutions designed to improve monitoring and contractor cooperation. The checklist can also be used to train new inspectors since it outlines not only the process of CQAP review but also the inspection process itself. In the long run, once the program has been proven to be statistically reliable and valid, it may be possible to include some factor representing a contractor's quality rating into the prequalification formula for future construction processes. In this way, contractors can be made accountable not only for completing a contract but also for providing a quality product and cooperating fully with Department personnel.

All of these uses, however, depend on proving the statistical validity of the review process and results. Whenever researchers or managers create a test or a measure of performance of any kind, several factors must be considered in evaluating how appropriate and useful the test is. These characteristics fall into two categories: those that are not statistical by nature, such as content or face validity, and those that are statistically based, such as reliability and statistical validity.

1. Face validity. Although validity is often a purely statistical term, there is a component of validity that is wholly intuitive and can be judged at the outset. Face validity is how valid or appropriate the test seems to the users and the developers. It includes the following:
   - Does the test or method seem to measure what one intends to measure (in this case, the control of the construction process and the contractor's implementation of the process)?
   - Does each item or section reflect the stated program objectives and does each comprehensively cover the activities one wishes to measure?
   - Will the rating of each item discriminate between projects that are "in control" versus those that are "out of control" (or will all projects score about the same)?
   - Is the vocabulary used in each item accurate and appropriate for the persons taking the test and are all items phrased in a readable and grammatically correct way?
   - Are all items of an appropriate difficulty? Is it too hard or too easy to meet the criteria?

For the most part, through development of the checklists, having them edited, going through the Delphi process, piloting the lists in the field, and gaining the backing of management, each of these questions has been answered.

2. Statistical validity. Statistical validity involves proving objectively that the test or measurement being used actually tests the behaviors or activities it is intended to measure. The key word is objectively. Measures of performance must be related to another objective indicator of quality: one that is well defined and has already been proven to measure construction quality. With regard to the CQAP checklists, there may be several outside measures one can use to determine validity. The question that may be asked is "Do projects that score high on the checklists produce a higher-quality product (or use a more efficient process) than those that score low?" To get at this, the question might be posed as "Do projects that score high on the reviews finish on time more often than those scoring low?" Program analysts might also determine if the project finished within budget or if it had fewer delays during construction. In the long run, analysts might examine the service life of the facility as a function of its CQAP score or use some other outside measure of quality (such as materials testing).

3. Reliability of the measure. Whereas validity of the measure is related to what is being measured, reliability relates to how it is being used. Reliability tests determine whether the same thing is being measured every time the scale is used or whether each person using it is actually measuring the same thing in the same way. If there are differences in reviewers' scores, it must be determined whether they reflect differences in the projects or just differences in the reviewers. Thus, this item is crucial.

The problem is that most measures of reliability are based on having different raters use the test to rate the same thing. If their ratings are different, this is because of the biases of the raters rather than differences in what they are rating. Unfortunately, the only way to obtain accurate reliability estimates is to have each reviewer rate the same projects and then compare their results. (There are measures of reliability

that compare even-numbered items with odd-numbered items on a checklist to determine internal reliability. These measures can be used to provide additional information, but they will not yield much insight into the interrater reliability problem.) Since reviewers are assigned to different geographic areas and do not rate the same projects, a clear measure of reliability has not been available. One possible solution to this problem would be to create a series of videotaped project reviews. All reviewers could then rate the various videotaped projects, and their results could be compared. Currently, the Council is considering the feasibility of such an effort.

4. Item analysis. In describing face validity (characteristic 1), some of the features discussed can be determined by simply looking at the checklists. However, there are statistical methods for objectively checking whether each of the qualities required under face validity is met. These statistical tests can be applied only when a sufficient body of data has been gathered. Once the data are available, the following indices can be generated:

- Difficulty index. The difficulty index indicates the average difficulty of each sublist and of the checklist as a whole. For all of the indices used in item analysis, the first step is to separate projects scoring high overall (the top third) from those scoring low overall (the bottom third). For each item, the proportion of each of the two most extreme groups answering the item correctly is calculated by dividing the number answering correctly by the total number in the group. The higher the difficulty index, the easier the checklist; the lower the index, the more difficult. Ideally, meeting the specifications listed in the checklist should not be too easy or too hard as measured by this index.
- Discrimination index. The discrimination index determines the extent to which the checklist (or checklist item) discriminates between projects that are in control versus those that are not (i.e., between those projects scoring high overall and those scoring low). This index is calculated by comparing the proportion of projects in the top group scoring high on an individual item with those in the low group also scoring high on that item. Clearly, if all projects, even those that are out of control, score high on an item, then the item does not discriminate between good and bad types of projects.
- Phi coefficient. The phi coefficient is a point biserial correlation between a high score on a checklist item and a high score overall. It is an index that measures internal consistency and indicates how well each checklist item predicts total review performance. Obviously, the more related an item is to overall performance, the more useful it is in discriminating between projects.

The idea behind using these indices in item analysis would be to develop the shortest and most concise checklist possible that would adequately discriminate between projects that are in control and those that are not. When one finds individual checklist items that have low discrimination indices, they would either be removed from the checklist or rephrased, possibly to be more specific concerning the activity the Department wishes to promote. When low phi coefficients are noted, it may be desirable to remove the item, move it to a checklist that is more closely related to it in content, or revise it to relate more clearly to the overall objectives of the particular checklist. The handling of items with overall high or low difficulty indices is trickier. If an item is high in difficulty but reflects the specification accurately, it would be illogical to remove or modify it (unless the wording was resulting in low scores). Perhaps contractor education on that particular item is needed. For low-difficulty items that reflect specifications that in other situations would be removed, no adjustment is needed since the purpose of the program is to encourage meeting the specifications.

Once these analyses are complete and the statistical validity of the CQAP has been documented, its results can be used to monitor the state of the construction process statewide and provide an additional tool to the Department to assist in creating better, more efficient, and longer-lasting facilities for the public.

REFERENCES

1. *Federal Aid Highway Program Manual*, Vol. 6, Chapter 4, Section 2, Subsection 8, Transmittal 372, April 2, 1982, p. 1–2.
2. I. J. Turshen and D. M. Miller. Study of Inspection In-Depth Program: Review of Regression Analyses and Report on the Quality Assurance Program. Department of Management Science, Virginia Commonwealth University, Richmond, 1986.
3. *Federal Aid Highway Program Manual*, Vol. 6, Chapter 4, Section 2, Subsection 8, Transmittal 372, April 2, 1982, p. 2.
4. M. G. Natrella. *Experimental Statistics*. National Bureau of Standards, U.S. Department of Commerce, 1966.
5. *Road and Bridge Specifications*, Virginia Department of Transportation, Richmond, 1987.
6. *Construction Phase Inspection Handbook*, Construction Division, Virginia Department of Transportation, Richmond, 1983.

The opinions, findings, and conclusions expressed in this report are those of the author and not necessarily those of the sponsoring agencies.

Publication of this paper sponsored by Committee on Quality Assurance and Acceptance Procedures.