GENERAL SERVICES ADMINISTRATION
Washington, DC 20405

July 8, 1994

FIRMR BULLETIN C-4
Revision 1

TO: Heads of Federal agencies
SUBJECT: Performance and capability validation of FIP systems

1. Purpose. This bulletin discusses factors affecting the selection and use of performance and capability validation techniques in acquiring Federal information processing (FIP) systems.
2. Expiration date. This bulletin contains information of a continuing nature and will remain in effect until canceled or superseded.
3. Contents.

Topic	Paragraph
Related material	4
Information and assistance	5
Definitions	6
Acronyms	7
Purpose of performance and capability validation	8
Validation costs	9
Diminished need for benchmarks and operational capability demonstrations	10
Projected workload and performance requirements	11
When compatibility-limited requirements apply	12
Assessing risks associated with incorrect sizing	13
Alternate validation techniques	14
Inspection of technical literature	14a
Rating charts	14b
Evaluator and peer experience	14c
Vendor evidence with agency validation	14d
Analytical modeling	14e
Simulation modeling	14f
Benchmarks	14g
Benchmarking with RTE	14h
Standard benchmarks	14i
Hybrid methods	15
Use of capability validation techniques	16
Cancellation	17
Outline of Performance and Capability Validation Planning for FIP Systems	Attachment A

4. Related material. Additional information on this subject may be found in:

FIRMR Section 201-20.304 - Capability and performance validation.

FIPS PUB 42-1, "Guidelines for Benchmarking ADP Systems in the Competitive Procurement Environment."

FIPS PUB 75, "Guidelines on Constructing Benchmarks for ADP System Acquisition."

GSA/KEES, "Use and Specifications of Remote Terminal Emulation in ADP System Acquisition."

FIPS PUB 101 "Guidelines for Lifecycle Validation, Verification, and Testing Computer Software."

NIST Special Publication 500-113, "Assessment of Techniques for Evaluating Computer Systems for Federal Agency Procurements."

NIST Special Publication 500-118, "A Guide to Performance Evaluation of Database Systems."

NIST Special Publication 500-123, "Guide on Information Workload Forecasting."

Federal Systems Integration and Management Center (FEDSIM) Publication, "Proceedings of the Symposium on Benchmarking and Alternatives," August 1989.

A Guide for Performance and Capability Validation.

5. Information and assistance. Additional guidance on information in this bulletin may be obtained from:

General Services Administration
Policy Analysis Division (KMP)
Washington, DC 20405

Telephone: FTS/Commercial (202) 501-2462

6. Definitions.

"Augmentation" means adding to or upgrading existing FIP hardware or software to increase its productivity or prolong its useful life.

"Capability validation" means the technical verification of the ability of a proposed FIP system configuration, replacement component, or the features or functions of its software, to satisfy functional requirements. The intent is to ensure that the proposed FIP resources can provide the required functions. FIP performance requirements are not implied or measured in the validation.

"Compatibility-limited requirement" means a statement of FIP resources requirements expressed in terms that require the items to be compatible with existing FIP resources.
"Performance validation" means the technical verification of the ability of a proposed FIP system configuration or replacement component to meet agency specified performance requirements.
"Price/performance" means the ratio of the price of acquiring a FIP system to that system's performance capabilities. In this ratio, performance may be expressed as the time required to perform a given workload.
"Validation budget" means the amount of resources an agency determines is appropriate for spending on performance and capability validation. The main factors influencing this amount are the risks associated with incorrect sizing and with acquiring a system that cannot perform required functions.

7. Acronyms.

CPU	Central Processing Unit
FIP	Federal Information Processing
FIPS PUB	Federal Information Processing Standards Publication
GAO	General Accounting Office
I/O	Input/Output
ITR	Internal Throughput Rate
NIST	National Institute of Standards and Technology
OCD	Operational Capability Demonstration
RTE	Remote Terminal Emulation
TPC	Transaction Processing Performance Council

8. Purpose of performance and capability validation. When acquiring FIP resources, agencies should ensure that the resources acquired will adequately fulfill the roles for which they are being acquired. The techniques used to obtain this assurance are referred to as either performance or capability validation. Performance validation measures the ability of a FIP system to meet agency specified performance requirements. It is generally associated with ensuring that the correct size of equipment is obtained. Capability validation is used to verify that an offering has a required capability. Attachment A outlines the steps for planning the validation.

9. Validation costs. A validation effort imposes a cost, both upon the agency and upon the offerors. An agency's objective should be to minimize the validation cost. The validation

selected should allow selecting officials to identify those offers that are capable of satisfying the agency's requirements.

10. Diminished need for benchmarks and operational capability demonstrations. In the early years of computing, comprehensive benchmarks, stress tests, and operational capability demonstrations (OCD's) were useful for validating reliability, performance and other requirements. In today's mature industry, the reliability and stability of the marketplace offerings are much higher. Also, there is substantial empirical data available from independent sources to assist agencies in assessing how an offering will perform in their environment and with their workloads. As a result, the use of benchmarks or OCD's may not be the most advantageous approach in many acquisitions. Agencies are advised to use the least costly technique that can satisfactorily validate offerings.

11. Projected workload and performance requirements. As part of the acquisition process, agencies must document in their requirements analysis the projected workload for the new system over its systems life (see FIRMR section 201-20.103-9). In developing its workload projection, an agency may use analytical techniques incorporating historical workload data from current systems, estimated data for new applications, and forecasts of workload changes over the system life, as appropriate. Workload may be described in terms of end-user functions, software activity, or hardware resources consumed. Performance requirements are generally expressed in terms of the time allowed for completing elements of workload. Typical examples include batch job throughput (jobs/time) and interactive terminal response time. Interactive response time requirements may be specified for various load conditions and also in terms of percentiles (e.g., the system should respond within 3 seconds 90% of the time for a type of transaction).

12. When compatibility-limited requirements apply. When the size of an installed base or the risks and costs associated with a software conversion justify a compatibility-limited acquisition, performance requirements may be specified using an internal throughput rate (ITR) (e.g., when acquiring mainframe processors). Performance requirements can also be expressed as a factor ("x" times) of a baseline system's performance for a given workload.

13. Assessing risks associated with incorrect sizing. The effects of acquiring insufficient capacity can range from a severe degradation in an agency's ability to perform its mission to a minor inconvenience. The effect of acquiring significantly more capacity than is needed can result in the Government investing more money over the life cycle than was actually needed. Agencies must assess the risks associated with incorrect sizing and consider these risks in establishing a validation budget. Where the risks of incorrect sizing are high, it makes sense to spend more on performance validation. When these risks are relatively low, an expensive validation effort is unwarranted.

14. Alternative validation techniques. A number of performance validation techniques are identified and their advantages and disadvantages are discussed in the following paragraphs.

Performance validation techniques include both manual and automated techniques that vary in expense, complexity, and reliability of results. The more rigorous techniques (i.e., benchmarking) require extensive computer resources to apply properly. All techniques require expertise. The value of performance validation depends upon the agency's ability to perform workload analyses that accurately measure current workloads and accurately forecast changes over the planned system life. The agency should choose the validation technique, or combination of technique(s), deemed most appropriate for the acquisition, considering cost and other factors discussed below.

a. Inspection of technical literature. Performance specifications validated by inspection of technical literature are sometimes appropriate. An acquisition of a small number of microcomputers might specify the microprocessor type and cycle time (e.g., a specified model with a specified speed). The agency would validate performance by inspection of a manual or product specification sheet.

b. Rating charts. These are commercially available computations (often in table form) of comparative information on the performance characteristics of different CPU's, disk devices, and other devices that are within vendor product lines and often, but not always, between vendors of compatible systems. By extrapolating performance on a current CPU, one can predict how an agency's workload will behave on a set of more powerful CPU's. However, since they only measure one aspect of computer performance, such as CPU speed, they may not be reliable indicators of overall performance, particularly if an agency's workload puts relatively more demands on other resources than CPU usage. Ways to reduce the risk of wrong sizing when using rating charts are discussed below.

(1) Types of rating charts. There are two types of rating charts - replicable and unreplicable. Replicable rating charts are empirically produced using publicly available benchmark tests. For example, there are charts based upon the Whetstone, Dhrystone and LINPACK benchmarks (these three products are discussed in paragraph 14i: Standard Benchmarks). Unreplicable rating charts are based on vendor claims, reports from users, or other sources for which a user cannot obtain sufficient data to reconstruct the test. Agencies that use unreplicable rating charts for specifying and validating capacity requirements should be aware that they may have difficulty defending their approach as equitable to all potential offerors.

(2) Correlation of agency workloads with rating charts. Agencies can increase their confidence in a replicable rating chart by correlating their workload to the workload used to produce the chart. Two approaches to correlation follow.

(i) Qualitative correlation. The agency can examine its workload characteristics and assess the extent to which they match characteristics of workload used to produce the rating chart(s). For example, an agency that processes almost exclusively a sequence of retrieve-update transactions may conclude that it correlates with the workload associated with a standard benchmark intended to measure performance with retrieve-update applications (see the discussion of the TPC-A benchmark in paragraph 14i below).

(ii) Quantitative correlation. The agency can construct its own benchmark, execute it on a sample of machines of different capacities, and mathematically correlate these results with the results shown on the rating chart. The effort involved may be similar to the effort of conventional benchmarking. This approach may be practical when the agency expects a large number of proposals and wants to avoid the time and expense of benchmarking each proposal.

(3) Additional considerations in the use of rating charts. These include the following.

(i) Many ratings are issued with a disclaimer about their accuracy. Agencies should consider these precautionary disclaimers in determining how much weight should be given to the ratings.

(ii) Ratings on newly announced computers may be less replicable than those for mature products, because the new equipment may not be readily available for testing. Reliance on such ratings should reflect this risk.

c. Evaluator and peer experience. This performance validation technique, based on rules of thumb or educated judgments, is easy to understand, quick and easy to use, and comparatively low in cost to the acquiring agency. It relies upon the technical judgment of the proposal evaluators gained through knowledge of, or direct experience with, the proposed FIP equipment components or configurations. Agency-verified FIP equipment performance rating charts may be an effective supplement for evaluator and peer experience in equipment acquisitions. Judgments may also be solicited from peers in the agency or elsewhere.

d. Vendor evidence with agency validation. In this method, commonly referred to as the "prove it" method, the agency describes its workload in the solicitation. Offeror proposals provide evidence of their system's performance. Such evidence may include rating charts, models, benchmark results, and other data. The type of evidence used may vary among offerors. In validating the offeror's proposal, an agency may supplement the evidence contained in the offeror's proposal with other information obtained from independent sources.

Agencies can use descriptions provided by the vendor to assess their confidence in the data and the risk of accepting it. Agencies should apply the same criteria to their independently-collected data as to the vendor-provided data.

e. Analytical modeling.

(1) Analytical modeling uses representations of the behavior of the components and processes of a computer system to predict its performance under varying workloads. Variables include the number of batch jobs or remote terminals, degree of multi-programming, and transaction volume and arrival rate. Queuing theory and other probabilistic techniques are often used. Approaches may include simple manual (pencil-and-paper) approximations, the use of general-purpose scientific programming languages (FORTRAN, PL/1, Pascal, etc.) for building the models, and the use of specialized analytical modeling languages (ACMS, BEST/1, ISS/THREE, MAP, RESQ, SCERTII, etc.).

(2) An advantage of analytical modeling is the insight it provides into performance of a system under changing workload conditions. For example, a multiprocessor's rating can be estimated using the rating of a "seed" processor and adjusting it for the changes resulting from connecting additional processors.

(3) Analytical modeling has several disadvantages as a validation technique. The models are often oversimplified in order to make them mathematically tractable, and this limits the inferences that may be drawn from them. The results are not often validated by measurement or simulation and, in cases where system evaluation studies have been carried out, the existing models have not seemed powerful enough to provide a uniform basis for measurement. Another disadvantage is that most of the literature on analytic modeling is a collection of analyses of specific models and, therefore, each new situation almost always requires a separate analysis by an expert.

f. Simulation modeling.

(1) Simulation modeling is frequently performed by using commercially available system simulation packages (such as GASP, GPSS, SIMSCRIPT, SIMULA, SLAM, etc.). Also, there are some packages that provide various functions such as predefined libraries of hardware and software performance characteristics, workload parameters defined by the user from historical accounting data, statistical subroutines, and pre-formatted reports. Simulation modeling may be used in many ways during the sizing and evaluation of proposed alternative equipment configurations. For instance, a model may be designed to simulate only the principal activities that occur within the computer as it operates or all the significant activities of the system. A model may also be implemented so that it is applicable to a one-of-a-kind special purpose system or may be general enough to represent an entire class of computers that includes many different manufacturers' systems.

(2) Simulation modeling can be highly accurate for comparisons of expected FIP equipment performance within a single manufacturer's line. Such accuracy, however, is dependent upon the model's accurate characterization of component performance.

(3) Simulation modeling is less accurate in a compatibility-limited architecture, and may have no validity in non-compatible architectural systems. However, when simulation models are properly developed and used, they can greatly reduce the agency's risk of acquiring inappropriate capacity during the evaluation and selection of new, replacement, or additional FIP equipment. In addition, the simulation packages can be used by the agency after the implementation of the selected equipment as part of the agency's capacity management program to predict the effects of system changes such as operating system enhancement, I/O peripheral upgrades or augmentations, and increased batch or on-line workloads.

(4) Data structured for simulation purposes should not be used as the only means of describing FIP resource requirements in solicitations. Simulation data should be accompanied by a narrative description of the FIP objectives and workload and also by application logic diagrams, if available.

(5) To promote competition, solicitations should not be structured in such a way as to require offerors to use a specific computer system simulator in order to submit their offers. A restrictive specification for a particular simulator is apt to hinder competition because some potential offerors may not want to incur the licensing cost for the prescribed product. Also, potential offerors may not want to invest in training their staff in how to use the prescribed product if they have no prior experience in using it. The acquiring agency should note, however, that if it allows each offeror to select its own simulator, comparing offers is apt to be more difficult. Similarly, the agency personnel may be forced into learning how to interpret the results of several simulator products. Agencies should balance the offsetting issues of greater competition versus internal efficiencies in evaluating the simulations.

(6) When offerors submit computer simulation as part of their offers, they should be required to describe clearly the simulation used and the make and model of the computer on which the simulation was run. The solicitation should also identify required simulation outputs, such as reports on major device utilization, response times and queue lengths.

g. Benchmarks.

(1) Benchmarks are specially constructed tests that verify the performance of a proposed FIP system by measuring its ability to execute within prescribed time limits a group of user programs representing a projected workload. The test specifications also assist offerors in judging the scale and complexity of equipment and software necessary to accommodate the user's workload. Generally, each responsive offeror must demonstrate the ability to run the tests successfully.

(2) Benchmarks are of two types: (i) natural benchmarks that employ the user's current program code to derive projected workload; and (ii) synthetic benchmarks that test hand-coded or automatically generated programs for system-to-system portability. The two types are often used together to create a more representa-tive set of workloads for the proposed system. Properly

constructed benchmarks will demonstrate that the offeror's proposed system contains adequate memory and I/O devices, the throughput speeds are sufficient to do the entire job, and the software proposed is operative and adequate. FIPS PUBS 42-1 and 75 provide detailed guidance on constructing and performing benchmark tests.

(3) While benchmarking is a rigorous technique, an improperly constructed benchmark, or a benchmark based on faulty workload projections, can yield a very precise but inaccurate result. The effective use of benchmarks is often inhibited by the agency's inability to develop tests that are portable across product lines without losing workload representation. Related to this is the need to ensure that the benchmark workload, usually a small sample of the projected workload, is not so small as to allow an offering to process it entirely within main memory and without realistic input/output processing. Another impediment might be determining the impact an offeror's modifications to the benchmark program has on system performance.

(4) Also useful for considering subsequent acquisitions within a given product line is a price/performance benchmark in which the cost of an existing machine or its successor is related to a specific performance standard.

h. Benchmarking with RTE.

(1) Agency requirements for large systems networks cannot easily be subjected to a benchmark test using the total proposed network of computers, terminal devices, and data communications facilities. RTE (remote terminal emulation) is a technique for conducting a benchmark test in such situations.

(2) RTE generally uses an external driver computer system to impose workload demands on the system under test. Potentially, many human-operator and remote device characteristics (e.g., interactive, transaction, and batch terminals) and actions can be represented precisely by the driver system in real time. The driver computer system can exchange control and application data transmissions with the system under test through that system's operational data communication hardware and software. RTE can use large numbers of data communication links of the same speeds, and with the same communication protocols, as in an operational environment.

(3) Another alternative is to use RTE with a synthetic benchmark taking the place of the external driver computer system. This approach can potentially save significant costs associated with the driver.

(4) When RTE is properly used, the system under test cannot distinguish whether a real or emulated device is generating the workload.

i. Standard benchmarks.

(1) Standard benchmarks are those benchmarks which have been developed by researchers, computer vendors, consultants or by the Government. These provide an objective measure of a system's performance at a relatively low cost (some may be free). Users of standard benchmarks should ensure that the standard benchmark's workload closely conforms to their own. Alternatively, users must have a reliable means to adjust the standard benchmark results to their own circumstances based upon the extent to which their own workloads may differ from those of the standard. A few of the more common benchmark programs or specifications are discussed below.

(2) Two synthetic benchmarks that have become industry standards for assessing computer processing performance are the Whetstone and Dhrystone programs. The Whetstone program, developed in 1964 at the United Kingdom's National Physics Laboratory, is intended to measure a computer's ability to process computational workloads. The Dhrystone program was first developed in 1984 and is designed to test a broader type of workload than is the Whetstone program. The Dhrystone program also contains a significant number of function and procedure calls.

(3) LINPACK is a collection of subroutines which solve systems of simultaneous linear equations. Developed at the Argonne National Laboratory in the mid 1970's, it has found use in measuring computer performance in solving dense systems of equations.

(4) The Transaction Processing Performance Council (TPC) had developed a number of standard benchmark specifications, each to be used for portraying a

particular type of application. The TPC's most robust specification, in terms of the number of transaction types included, is TPC-C. This standard specification, issued in July 1992, models an order entry workload.

(5) In determining whether or not an agency can effectively use a standard benchmark, it must evaluate how closely its unique workload compares to the workload used to establish the standard benchmark. In some instances, an agency's projected workload will not exactly conform solely to a standard workload associated with a standard benchmark. Rather, it may involve a projected workload that has elements constituting the essential activity of a number of different standard benchmarks. In this case the agency may be able to estimate roughly for the various offerings, their likely performance with regard to those aspects of its forecasted workload that conform to each relevant standard. The agency could then weigh the results with regard to each relevant standard accordingly.

15. Hybrid methods. Agencies may combine several performance validation methods when the combination reduces the risk of inappropriate sizing or cost to offerors and/or the government, or enhances competition. For example, a benchmark may test the CPU and I/O subsystems while performance specifications are used for the modems (i.e., 9600 baud).

16. Use of capability validation techniques.

a. Capability validation techniques verify the proposed FIP system's technical ability to satisfy purely functional requirements specified in the solicitation or in the manufacturer's technical literature. The primary purpose is to ensure that any items to be acquired can successfully perform the specified or proposed functions. As a validation tool, they are not intended to measure and verify performance.

b. Capability validation is usually accomplished by examining the technical specifications and associated technical literature. In some cases, specific technical questions regarding particular points may be posed to the offeror or to users of the proposed system. In still other cases, operational capability demonstrations (OCD's) may be required. OCD's are functional demonstrations that the item to be acquired can perform the required function. All OCD's should be designed and conducted to prove that the proposed items meet the capabilities described in

the solicitation, and that they operate effectively as part of an integral FIP equipment system. During the OCD, no system performance requirement should be imposed or measured by the agency. However, OCD's do provide opportunities to observe other characteristics, such as ease-of-use factors. Since validation of functional capabilities cannot serve as a representative workload test of performance, OCD's may supplement but can not replace any technique for validating performance. Validation requirements should be considered when selecting a capability validation technique.

17. Cancellation. FIRMR Bulletin C-4 is canceled.

JOE M. THOMPSON
Commissioner
Information Resources
Management Services

FIRMR Bulletin C-4
Revision 1
Attachment A

OUTLINE OF PERFORMANCE AND CAPABILITY VALIDATION PLANNING FOR FIP SYSTEMS

- o Analyze the Workload
  - Current workload
  - Projected workload changes over the system life

- o Determine System Validation Requirements
  - Performance requirements
  - Capability requirements

- o Identify Alternative Validation Techniques
  - Performance validation techniques
    - --Inspection of technical literature
    - --Rating charts
    - --Evaluation and peer experience
    - --Vendor evidence with agency validation
    - --Analytic modeling
    - --Simulation modeling
    - --Hybrid methods
    - --Benchmarking (including remote terminal emulation and the use of standard benchmarks)
  - Capability validation techniques
    - --Inspection of technical literature
    - --Operational capability demonstration

- o Analyze the Risks
  - Strengths and weaknesses of validation techniques
  - Impact of mission disruption
  - Impact of system life cost

- o Select Appropriate Validation Technique(s)
  - Protection against adverse impact on agency mission
  - Cost-effectiveness

- o Incorporate Validation into Appropriate Acquisition Plan Phases
  - Requirements definition/workload analysis
  - Specification preparation
  - Preaward proposal evaluation
  - Postaward acceptance testing
  - Option executions