

Evaluation of the Judicial Performance Program

Prepared for:

**Chief Justice Ronald T.Y. Moon of the Supreme Court
The Judiciary
State of Hawaii**

Prepared by:

Michael Heim, Principal Investigator

Dr. Glenn Hirata, Ph.D., Evaluation Specialist

March 1995

Evaluation Summary

Background. The present report documents an external evaluation of Hawaii's relatively new and still evolving Judicial Performance Program. A combination of circumstances, particularly the accumulation of sufficient data that could now allow the preparation of judicial profiles for Circuit Court judges in the First Judicial Circuit, and the phased-in expansion of the program to the District Court, followed by Family Court of the First Circuit and all Neighbor Island circuits, make the present time a uniquely opportune one for the kind of studied reflection that evaluation can provide. The validity of an instrument, like the *Lawyer's Questionnaire*, is intimately dependent upon the interpretation and uses of the data derived from it. Considering that the Judicial Performance Program may soon be capable of producing judicial profiles for judges assigned to Circuit, District, and Family Courts in all circuits statewide, basic questions about proper interpretation and appropriate uses of the data are best addressed now.

Purpose. The scope and purpose of the evaluation were shaped by the contractual relationship between the evaluators (Contractor) and The Judiciary, and more particularly by The Judiciary's specifications for the evaluation. The design for the current work was outlined in our proposal prepared in response to The Judiciary's Request for Proposal (RFP) No. J95091. In brief, the RFP specified that the proposed evaluation address three objectives, which were included in our proposal and defined the purposes of the evaluation:

- Assess instrument validity;
- Assess validity of evaluation procedures; and,
- Assist with data interpretation and presentation.

Methods. Several methods were used for assessing instrument validity and for assessing the validity of the evaluation procedures used in the Judicial Performance Program. For assessing instrument validity, we conducted analyses of *Lawyer's Questionnaire* data from the Circuit Court and the District Court of the First Circuit. The data were used to estimate the reliability of ratings, amount of sampling error in mean scores, and investigate several issues related to appropriate data interpretation. A content analysis of the instrument was also conducted in which the Questionnaire's organization, format, and type of items used were examined. In addition, a series of telephone interviews were conducted with administrative and senior judges as well as selected members of the Special Committee on Judicial Performance. The interviews collected information relevant to key issues of instrument validity and judicial evaluation procedures.

The validity of evaluation procedures used in the Judicial Performance Program was assessed via four methods. A review of the literature was conducted that provided us with substantive background about the current state of the art in judicial performance evaluation, and indications of problems or questions about evaluation processes or the uses of judicial performance data that might be applicable to the present study. Next, comprehensive reviews of the Judicial Performance Program were conducted using two independently developed evaluative frameworks: (1) the Personnel Evaluation Standards, which are intended for use in assessing the adequacy of personnel evaluation systems, and (2) the American Bar

Association's Guidelines for the Evaluation of Judicial Performance. In addition, we conducted a "walk through" of actual program operations by Planning and Evaluation Division staff that served partly as on-site verification of information previously collected from documents and informal interviews, and as a first-hand inspection of confidentiality measures, data handling and analysis routines.

Conclusions and Recommendations

Reliability estimates (Cronbach's alpha coefficient) were .97, .94, and .96 for the Legal Ability, Judicial Management Skills, and Comportment scales, respectively. Limitations on the precision of mean scores per judge are due largely to sampling error, that is, the sampling variation introduced via the particular set of attorneys who happen to be surveyed. As a rough rule-of-thumb, means for each of the three scales would have to be different by at least 0.5 of a scale point (e.g., half the distance between "Good" and "Excellent" on the rating scale), for samples of 20 respondents, before one might conclude that a statistically significant mean difference had been found.

Several questionnaire items were found that had relatively large rates of 'Not Applicable' (NA) response. Motions, as contrasted with trials, incurred higher rates of NA response, suggesting that the applicability of the questionnaire's content may be limited by the type of proceeding. Generally, it did not appear that the occurrence of NA responses and omitted items (blanks) was associated with any systematic increase or decrease in mean ratings for individual judges.

Interviews conducted with administrative and senior judges as well as members of the Special Committee on Judicial Performance clearly indicated that the Judicial Performance Program does *not* negatively interfere with normal courtroom practices, nor is it perceived currently as infringing upon the independence or integrity of the judiciary. Most interviewees reported that the *Lawyer's Questionnaire* covered almost completely the duties and responsibilities of a judge; however, there were some questions regarding kinds of proceedings, and judges, for which the instrument is appropriate.

The Judicial Performance Program's evaluation process was found to be basically sound according to Personnel Evaluation Standards and the American Bar Association's Guidelines, when the assumed purpose was that of improving individual judges' performances through evaluative feedback to the Chief Justice and to participating judges. The Program held up quite well against the general, exemplary requirements of the Standards and the ABA Guidelines which are more specific to the judiciary. Importantly, the ratings on the Standards were strong in areas such as measurement reliability, validity and evaluator credibility.

One area of some uncertainty is that of case selection. It is not known at the present time whether potential conflict of interest is problematic in the process of court staff determining which cases constituted "meaningful opportunities."

The ultimate appraisal of the validity of the Program's evaluation process rests heavily on the eventual use and dissemination of performance evaluation results. It appears to have been a

good move to have piloted and phased-in the early implementation of the Program. The decision to defer reporting of performance evaluation results for individual judges until the soundness of the Program overall could be assessed also seems to be a wise decision.

The Program's administration and support appears to be quite adequate. There is good overall monitoring of data collection and procedures. Access to files is adequately controlled and confidentiality is not compromised. What the Program does seem to lack is an overall long-range plan.

The following recommendations have been selected from the full body of the present report:

- Examine more carefully the present procedure used for case selection. Specifically, the operational definition of "meaningful opportunity" needs improvement, possibly by considering a dual requirement of substantive content and a minimum time duration.
- Continue to keep case and court identifying information (Judicial Survey Information Form) separate from the *Lawyer's Questionnaire*. Current practices serve to protect the confidentiality of respondents and should be continued.
- Narrative comments that respondents may supply should be "sanitized" to remove any personally identifying information, and provided *only* in profiles that are distributed to the judge involved and the Chief Justice. Such comments should not be included in other reports.
- Consider developing and testing a "high volume and special proceedings" lawyer's questionnaire for District Court proceedings.
- An aggregate summary that could be used to provide a snapshot of judges' performance collectively is recommended. The summary could be used for designing judicial education activities, and, possibly, might be used for public release.
- Profiles for self-improvement use by individual judges should provide descriptive statistics for items and for each of the three scales (Legal Ability, Judicial Management Skills, and Comportment).
- Study carefully the release of results for secondary purposes such as retention. There is potential here for conflict with the purposes of self-improvement.
- Consider issuing a policy statement regarding the sanctioned uses and dissemination of Program results.

Table of Contents

Evaluation Summary	i
I. Introduction	
A. Overview of Hawaii's Judicial Performance Program	1
B. Scope and Purpose of the Evaluation	3
II. Methods	
A. Procedures for Assessing Instrument Validity	5
(1) Data Analysis	5
(2) Content Analysis	6
(3) Interviews	7
B. Procedures for Assessing Validity of the Evaluation Process	8
(1) Literature Review	8
(2) Comparisons with Personnel Evaluation Standards	9
(3) Comparisons with American Bar Association Guidelines	10
III. Findings	
A. Instrument Validity	11
(1) Results of Data Analysis	11
(2) Results of Content Analysis	27
(3) Results of Interviews	30
B. Validity of the Evaluation Process	34
(1) Results of Personnel Evaluation Standards Comparisons	34
(2) Results of American Bar Association Guidelines Comparisons	40
(3) Correspondence between Standards and Guidelines	47
C. Data Interpretation and Presentation	49
IV. Conclusions	50
V. Recommendations	52
References	54
Appendices	
A: <i>Lawyer's Questionnaire</i> (sample copy)	
B: Additional Data Analysis Summary	
C: Interview Protocols	
D: Interview Data Analysis	
E: Personnel Evaluation Standards (adapted to the judiciary)	
F: Sample Ratings Worksheet, Personnel Evaluation Standards (adapted to the judiciary)	
G: Correspondence between ABA Guidelines (adapted in checklist form) and the Personnel Evaluation Standards	

I. Introduction

The present report documents an external evaluation of Hawaii's relatively new and still evolving Judicial Performance Program. A combination of circumstances, particularly the accumulation of sufficient data that could now allow the preparation of judicial profiles for Circuit Court judges in the First Judicial Circuit, and the phased-in expansion of the program to the District Court (beginning in late 1993) followed by Family Court of the First Circuit and all Neighbor Island circuits (Circuit, District, and Family Courts beginning in late 1994), make the present time a uniquely opportune one for the kind of studied reflection that evaluation can provide. The evaluation focused on two key issues of validity that are critical at the present time: the content validity of the *Lawyer's Questionnaire* instrument and the validity of the associated judicial evaluation process. The validity of an instrument, like the *Lawyer's Questionnaire*, is intimately dependent upon the interpretation and uses of the data derived from it. Considering that the Judicial Performance Program may soon be capable of producing judicial profiles for all Circuit, District, and Family Courts in all circuits statewide, basic questions about proper interpretation and appropriate uses of the data are best addressed now.

The present report is organized into five major sections with several sub-sections each. In the remainder of the current "Introduction" we present a sub-section that contains a brief overview of Hawaii's program, followed by a sub-section that delineates the scope and purpose of the external evaluation. The next major section, "Methods," is organized into two parts, which describe the procedures used in the current evaluation to assess instrument validity and the validity of the evaluation process. "Findings" are next presented in three subsections, "Instrument Validity," "Validity of the Evaluation Process," and "Data Interpretation and Presentation." The report concludes with "Conclusions" and "Recommendations" as the fourth and fifth major sections. References and appendices containing adjunct materials are also provided.

A. Overview of Hawaii's Judicial Performance Program

Within the last two decades, judicial performance evaluation programs have been implemented or are currently under active development in about twenty states. While these programs have some similarities, "each is designed to accomplish goals suited to and established by the individual jurisdictions" (Keilitz & McBride, 1992, p. 4). Thus, for instance, Alaska's program, the longest established (authorized by the legislature in 1976), is designed to provide recommendations to voters in judicial retention elections. By contrast, New Jersey's court-sponsored program which began as a pilot program in 1983 and was made permanent in 1986, and on which Hawaii's program has been largely modelled, is designed primarily to provide systematic information feedback to judges to improve judicial performance.

While the goal of improving judicial performance individually and institutionally is important, the means by which it is done is equally important. Clearly, as has been evident in the development of the Judicial Performance Program in Hawaii like judicial evaluation programs

elsewhere, cautious small-scale piloting and careful phased-in implementation are essential to ensure that other compelling needs are also met: protecting the independence and integrity of the judiciary, and preserving the anonymity of respondents. The two go together in this context.

Below is a "thumb-nail" chronology of the development of Hawaii's Judicial Performance Program.

What?	When?	Synopsis
Idea introduced	Mar. 1985	Chief Justice Herman Lum introduced the idea of evaluating trial judges for purposes of improving judges' performance.
Concept supported	Nov. 1986	The Committee on Judicial Evaluation presented a report at the Hawaii State Judicial Conference advocating a judicial evaluation program.
Authorization	Nov. 1990	Supreme Court Rule 19 formally established the Judicial Performance Program (effective 1/91).
Administration	Feb. 1991	The 13-member Special Committee on Judicial Performance was appointed to provide guidance and oversee the program's administration and development.
Test Phase	July 1991- Dec. 1991	"Test Phase 1" -- First Circuit Court, Criminal Division; four (4) trial and motions judges; 120 questionnaires distributed.
	Jan. 1992- July 1992	"Test Phase 2a" -- First Circuit Court, Criminal Division; 17 Criminal Division trial and motions judges; 340 questionnaires mailed to attorneys.
	Aug. 1992- Jan. 1993	"Test Phase 2b" -- First Circuit Court, Civil Division; nine (9) trial and motions judges; 232 questionnaires distributed.
	May 1993- June 1993	"Test Phase 3" -- First District Court, Civil, Criminal, DUI, Traffic Divisions; five (5) trial and motions judges; 238 questionnaires distributed to attorneys and <i>pro se</i> litigants.
Permanent JPP	Aug. 1993	Chief Justice Ronald T.Y. Moon announced that the program was permanent and would be expanded to all judges and justices.
	Sept. 1993- Mar. 1994	Inauguration of permanent program with the Circuit Court of the First Circuit; 10 judges with 2+ years bench experience; 284 surveys distributed by court staff.
	Nov. 1993-	Inauguration of the permanent program with the District Court of the First Circuit; 11 judges with 1+ years bench experience; 136 surveys distributed by court staff as of 3/22/94.

By the end of 1994, initial implementation of the permanent Judicial Performance Program had expanded to include Family Court and all circuits statewide.

The stated objectives of Hawaii's Judicial Performance Program are as follows (verbatim from *Report of the Judicial Performance Program*, p. 4):

1. Improve individual judges' performances by providing information to the Chief Justice;
2. Provide the Judicial Selection Committee with a potential source of information for retention decisions;
3. Facilitate the Chief Justice's effective assignment and use of judges;
4. Improve the design and content of judicial education programs; and
5. Assist the Chief Justice to discharge administrative responsibilities.

As noted in the preceding synopsis, the program is sponsored by the Supreme Court of the State of Hawaii and is provided overall guidance and administrative oversight by a 13-member Special Committee on Judicial Performance. The Special Committee is comprised of three (3) non-lawyers, the (1) Administrative Director of the Courts, six (6) members of the bar, and three (3) judges. Staff support is provided by the Planning and Evaluation Division located administratively within the Office of the Administrative Director of the Courts.

The evaluation data used in the program is based entirely on the *Lawyer's Questionnaire* (see *Appendix A* for a sample copy), an instrument adapted from a questionnaire with the same title used in the New Jersey program. Content of the questionnaire and procedures associated with its use will be described in considerable detail later in the present report.

B. Scope and Purpose of the Evaluation

The scope and purpose of the evaluation are shaped by the contractual relationship between the evaluators (Contractor) and The Judiciary, and more particularly by The Judiciary's specifications for the evaluation. The design for the current work was substantially presented in our "Proposal to Review and Evaluate the Judicial Performance Program's Instrument, Methodology, and Processes for the Chief Justice" (dated October 26, 1994) which was prepared in response to The Judiciary's Request for Proposal (RFP) No. J95091.

In brief, the RFP specified that the proposed evaluation address three objectives, which were included in our proposal, and which essentially defined the purposes of the evaluation:

- Assess instrument validity;
- Assess validity of evaluation procedures; and,
- Assist with data interpretation and presentation.

The specific methods used to achieve these objectives are described in the following section of this report. For now it should be mentioned that the scope of work for the first objective, "Assess instrument validity," was necessarily limited to: (1) analysis of the psychometric

(measurement) quality of the *Lawyer's Questionnaire*, with specific emphasis on the reliability of the scores/ratings produced; and (2) an assessment of the instrument's content validity. Given the limitation of no available empirical data other than that from the questionnaire, it was not possible, for example, to examine the instrument's concurrent or predictive validity.

II. Methods

A. Procedures for Assessing Instrument Validity

(1) *Data Analysis.*

A series of data files containing Judicial Performance Program questionnaire records were made available for analysis in the current evaluation. It should be noted that none of the files contained data elements that would have allowed us to determine the identities of the judges rated or of the lawyer-respondents involved; anonymity of the participants in the Judicial Performance Program was preserved. The primary set of analyses were conducted on two versions of questionnaire data for the Circuit Court of the First Judicial Circuit. The first file of Circuit Court data, referenced herein as "Duplicated Data File," contained 311 records for 12 judges. The file is referred to as "duplicated" because it contained instances whereby two or more records were from the same attorney-respondent completed for the same judge (but from different proceedings). It is the current intent of Planning and Evaluation Division staff to eliminate duplicated records prior to generating judicial profiles.

The second data file was a subset of the first data file. It contained no "duplicated" records, in the sense defined above, and is referred to herein as the "Unduplicated Data File." This file contained 287 records. Analysis of the two Circuit Court data files was conducted in parallel fashion because one of the questions of interest was to what extent the two files would generate similar (or different) findings. While, perhaps, seemingly tangential to the purpose of assessing instrument validity, findings about the extent of similarity or differences between the two ways of analyzing the available data are of direct operational interest to Planning and Evaluation Division staff and, potentially, could have validity implications for judicial profile data interpretation and use.

With the exception of a few computations more readily done manually, nearly all data analysis was conducted using Statistical Analysis System (SAS Institute Inc.) software on an IBM-compatible PC. While the reliability coefficients and estimates of measurement and sampling error obtained from the questionnaire data analysis are of central interest, analysis was also conducted to generate various descriptive statistics per item and scale, to examine variation in mean scores of ratings for judges, and to explore relationships of several selected variables with mean scores (e.g., the relationship between attorney type, whether prosecutor/plaintiff or defense, and mean scores). Some additional analyses simply described available "background" variables such as type of proceeding (civil or criminal), number of times the respondent had appeared before a given judge, and so on. Results of data analysis peripheral to issues of reliability and validity are reported in *Appendix B*, rather than in the main body of the report.

Finally, a secondary set of analyses were conducted on available District Court of the First Circuit data from 116 respondents. While an unduplicated version of these data was available ($n = 91$), the file for the 116-respondent "duplicated" version was used because of its somewhat larger size and the increased likelihood that estimates obtained from it would be

more statistically stable. This data file was analyzed in a confirmatory mode, that is, we attempted to replicate several key findings from the Circuit Court data analysis with the District Court data.

(2) Content Analysis.

In the context of the *Lawyer's Questionnaire*, content validity refers to the extent to which the instrument's items adequately sample or represent the constructs implied by the instrument's three sections or scales: "Legal Ability," "Judicial Management Skills," and "Comportment." Some indirect statistical evidence can be used to investigate content validity. At least a moderate level of internal consistency among items, for instance, would be expected. (Internal consistency is indexed by the type of reliability coefficients that were generated for this evaluation.) While statistical evidence can be partially helpful, "inevitably content validity rests mainly on appeals to reason regarding the adequacy with which important content has been sampled and on the adequacy with which the content has been cast in the form of test items" (Nunnally, 1978, p. 93).

Beyond the inspection of psychometric properties associated with the *Lawyer's Questionnaire*, other approaches used to assess the instrument's content validity included analyses of:

- (a) coverage of major duties and responsibilities of judicial performance;
- (b) scaling characteristics and anchor descriptors; and
- (c) ancillary items soliciting information on respondents' demographic characteristics and comments.

These analyses reflect the multi-faceted approach required to assess content validity. For the purposes of this evaluation, assessments of the extent and adequacy of the *Lawyer's Questionnaire* in covering the major duties and responsibilities expected of judges were based both on appraisals by administrative and senior judges (through telephone interviews) who participated in the Judicial Performance Program, and on evaluation performance criteria cited in the American Bar Association Guidelines (ABA, 1985). For the sake of brevity, results from these appraisals and performance criteria will be discussed under the forthcoming sections, "Interviews," and "Comparisons with ABA Guidelines."

The areas of primary interest, here in this content analysis section, were limited to that of examining the instrument's organization, format, and type of items used. This included an examination of the *Lawyer's Questionnaire* scale and anchor descriptors for the rating categories, first independently, and then in comparison to similar instruments developed elsewhere. Also, the additional items in the "Background Characteristics" and "Comments" sections were reviewed in terms of clarity and utility.

(3) Interviews.

A series of telephone interviews was conducted by the evaluators with two sets of target respondents or interviewees. The first set of target interviewees included ten (10) administrative judges and senior judges, all of whom were participants in the Judicial Performance Program. The judges represented all judicial circuits statewide and the Circuit, District, and Family courts. The second set of target interviewees included eight (8) members of the Special Committee on Judicial Performance; however, two (2) of these Committee members were also among the ten judges in the first group of interviews. With the exception of the two judges who belonged to both groups, then, six (6) other members of the Special Committee were actually selected. The members of the Special Committee selected included the Chairperson, Vice-Chairperson, two judges, two attorneys, and two non-attorneys. (Neither time nor other available evaluation resources permitted us to conduct the interviews as personal interviews, or to expand the scope of the interviews to include more respondents.)

The interviews provided an excellent opportunity to retrieve information directly from the program's participants (judges) and the program's developers/planners (Special Committee members) for several questions we had developed. The focus of our questions concerned issues of instrument validity as well as evaluation process. (Copies of the interview protocols used are given in *Appendix C*. The Appendix also contains a listing of the target interviewees.)

All interviews were conducted during the period January 20-31, 1995. Of the 16 individual interviewees targeted, interviews were completed with all (100%).

Analysis of the interview data used primarily the qualitative method of content categorization to summarize narrative responses, followed by tabulation to obtain the counts of responses within categories. For most interview questions, then, the findings were summarized in the form of tables showing either the number of respondents, or the number of "mentions," for each category of narrative response provided. (Open-ended questions or requests for comments or elaboration often result in respondents providing several responses or "mentions" spanning different categories. For instance, a given interviewee might generate three different mentions in response to a single follow-up question.) The full summary of the interview analysis is given in *Appendix D*. Selected highlights are given in the main body of the present report.

B. Procedures for Assessing Validity of the Evaluation Process

(1) *Literature Review.*

The general purpose of the literature review was to develop a good understanding of the purposes, state of development, progress, and possible issues with judicial performance programs, especially as related to the evaluation processes and instruments used therein. More specifically, the literature review served as essential preparation for the present evaluation by providing us with: (1) substantive background information about the current state of the art in judicial performance evaluation; and (2) indications of problems or questions about evaluation processes or the uses of judicial performance data that might be applicable to our present study of Hawaii's Judicial Performance Program.

Background information about judicial performance evaluation programs in other jurisdictions, particularly information about the evaluation methods (processes and instruments) used in those programs, was obtained in three ways. First, and most productive in terms of yielding the largest number of relevant documents, we obtained copies of many useful documents directly from our Planning and Evaluation Division liaison for the present evaluation. Included, as examples, were documents such as the *American Bar Association Guidelines for the Evaluation of Judicial Performance* (ABA, 1985); various memoranda from the National Center for State Courts which contained informative appendices containing lists of contact persons and brief descriptions of judicial performance programs across the nation; and, reports from the judicial performance programs in New Jersey and Connecticut.

Second, we conducted electronic searches -- via the UHCARL Library System at the University of Hawaii, Manoa -- of the University of Hawaii libraries, particularly the William S. Richardson School of Law Library, and other selected law libraries on the mainland (e.g., The Northeastern University Libraries Information System, Boston, MA). The electronic searches were successful in identifying published documents such as the ABA's *Guidelines* and relevant journal articles (e.g., papers published in the *State Court Journal*). It appears, however, that such searches, which tap into the on-line catalogues of university library collections, do not usually access either the kind or level of detailed information about judicial performance programs that we wanted.

Third, given the adaptation of Hawaii's program based on that of New Jersey, we contacted Mr. Richard Young, Assistant Project Director, Supreme Court of New Jersey. Our inquiries of Mr. Young focused on two areas: additional information about some procedural details of New Jersey's program that have been adopted in Hawaii's program (e.g., the recent change from the use of "matched" to "unmatched" questionnaires) and the availability of additional validity research or evaluation information. In addition, we also contacted Ms. Margery M. Wilbur, Judicial Evaluation Administrator, Office of the Chief Court Administrator, Hartford, Connecticut. Ms. Wilbur provided us with updated information about Connecticut's Judicial Performance Evaluation Program.

(2) Comparisons with Personnel Evaluation Standards.

Central to an appraisal of the Judicial Performance Program's evaluation process is the application of 21 Personnel Evaluation Standards. These Standards help to assess the soundness of personnel evaluation *systems* and require that evaluations be proper, useful, feasible, and accurate. Developed by the Joint Committee on Standards for Educational Evaluation with representatives from 14 prominent national associations concerned with education, measurement, and evaluation, the Standards collectively are comprehensive and general criteria by which institutional policies and procedures for evaluating personnel may be developed, reviewed, upgraded or implemented.

The development of the Standards themselves underwent an impressive series of consensus checks and validation processes. First, the Joint Committee investigated personnel evaluation practices and obtained input from hundreds involved with the evaluation of personnel. Next, a national panel of writers helped draft the standards. Then, several national and international Review Panels critiqued the draft which was subsequently revised by members on the Joint Committee. Another 40 professionals helped to critique the second draft in public hearings. This second draft was also field-tested in a number of institutional settings. Throughout this process, a Validation Panel helped to monitor and evaluate the project overall.

The content and essence of the Standards have broad application to many professions. Here, they are adapted to the judicial field. It is important to note that the appropriateness of applying any of the 21 Personnel Evaluation Standards is dependent on the specific purpose or function of interest. Here, the function of interest is assumed to be that of performance review in general rather than that, for example, of certifying, licensing, or terminating. With regard to the Judicial Performance Program, the function of interest translates into the Program's first and primary objective to "Improve individual judges' performance by providing information to the Chief Justice" (*Report of the Judicial Performance Program*, p. 4).

Appendix E is a list of all 21 Standards, with descriptions adapted to the judicial field. The Standards are organized into the four broad domains of propriety, utility, feasibility, and accuracy.

For purposes of the present evaluation, a worksheet rating form was then developed for use by the two evaluators (*Appendix F*). Each evaluator then independently applied all 21 Standards in appraising the extent to which the Program does or does not meet procedural requirements or performance criteria specified in the Standards. [A 3-point scale was used to rate each Standard (Met=3; Partially Met = 2; Not Met = 1).] The evaluators then reviewed the initial ratings for all 21 Standards, discussed common findings and discrepancies, and reached consensus on final ratings.

(3) *Comparisons with American Bar Association Guidelines.*

Another key source of information useful for purposes of performance review, specific to the judicial field, is the series of guidelines published by the ABA. Together with the Joint Committee's Standards, these Guidelines serve as a foundation upon which the Judicial Performance Program's policies, procedures, criteria, and instruments can be systematically reviewed. The Standards and the ABA Guidelines are complementary. The Standards are general in its purview, wide in its scope. The ABA Guidelines, on the other hand, are specific to the judiciary, and directly applicable to duties and responsibilities of judges.

The *American Bar Association Guidelines for the Evaluation of Judicial Performance* (ABA, 1985), approved by the ABA's policy-making House of Delegates in July 1985, is intended as a "rough template, a checklist, a series of guidelines which should serve jurisdictions that wish to engage in judicial evaluation" (p. ii). The scope of the ABA Guidelines comprises five parts: "Goals and Uses," "Administration and Support," "Criteria," "Methodology," and "Uses and Dissemination." Labels for two of these categories were slightly modified in what follows. "Goals and Uses" was changed to "Goals and Purposes," and "Criteria" was expanded to "Performance Criteria."

An approach similar to that of applying the Standards was developed for the ABA Guidelines. However, in adapting the Guidelines, it was first necessary to develop a working checklist with items that each delineated a single recommended procedure or criteria (some of the Guidelines are composed of dual procedures or multiple criteria). In so doing, we developed a 47-item checklist addressing program policy, procedural requirements, and performance criteria recommended for the evaluation of personnel performance (see *Appendix G*).

As with the Standards, the two evaluators first independently assessed the Judicial Performance Program with respect to the Guidelines checklist, and rated the extent to which the requirements or criteria were being met. The same 3-point scale used for the Standards (Met, Partially Met, Not Met) was used for the Guidelines. Then, the evaluators reviewed the initial ratings, discussed common findings, and discrepancies, and reached consensus on final ratings. As will be noted in the next section, there were a number of checklist items not possible to rate at this time, typically because the item had not been implemented as yet.

III. Findings

A. Instrument Validity

(1) *Results of Data Analysis.*

All findings in this section are from analyses of Circuit Court of the First Circuit data unless noted otherwise.

Two data files were provided by The Judiciary, Office of the Administrative Director, Planning and Evaluation Division for use in the current evaluation. The first data file (je4.d, dated 12/19/94), referenced herein as "Duplicated Data File," contained 311 Judicial Performance Program questionnaire records for 12 judges from the Circuit Court of the First Judicial Circuit. The file is referred to as "duplicated" because it contained instances whereby two or more records were from the same attorney-respondent completed for the same judge (but from different proceedings). The second data file (je4.e, dated 12/20/94) was a subset of the first data file. It contained no duplicated records, in the sense defined above, and is referred to herein as the "Unduplicated Data File." This file contained 287 records. Both data files were from the current implementation of the permanent Judicial Performance Program.

Results of the analyses are organized immediately below into the following sections:

(a) Questionnaire Item Statistics, (b) Questionnaire Scale Statistics, (c) Results by Judge, (d) Exploration of Relationships, and (e) Partial Confirmation with District Court Data.

(a) Questionnaire Item Statistics

Both the frequency distributions of item option selection and summary descriptive statistics were computed. The three-part table that follows is a composite drawn from both types of analyses. (We do not report the frequency distributions here simply because of their cumulative length.) Most notable in the table that follows (Table 1) is occurrence of markedly large numbers of 'Not Applicable' responses to some items (e.g., Scale 1, Items # 3, 8, or 11). More will be said of this later. Also, it might be noted that the item statistics from the Duplicated and Unduplicated data files are quite similar.

Table 1. Panel (a). *Lawyer's Questionnaire* item statistics.

Scale 1: Legal Ability					
Item (paraphrased)	Mean	S.D.	N	'Not Applicable' (#)	missing (#)
1. know relevant law	4.1	1.03	292	13	6
	4.1	1.05	268	13	6
2. know rules & procedures	4.3	0.85	284	23	4
	4.3	0.86	261	22	4
3. know rules of evidence	4.2	0.90	206	100	5
	4.2	0.92	189	93	5
4. ability to identify issues	4.1	1.02	297	7	7
	4.1	1.04	274	6	7
5. judgement applying law	3.9	1.20	291	11	9
	3.9	1.23	268	10	9
6. giving reasons for rulings	3.9	1.12	277	31	3
	3.9	1.14	254	30	3
7. clarity of explanations	3.9	1.12	283	25	3
	3.9	1.14	260	24	3
8. adequacy of findings	3.9	1.06	175	129	7
	3.9	1.08	158	122	7
9. clarity of decision	4.0	1.06	267	40	4
	4.0	1.08	244	39	4
10. completeness of decision	3.9	1.07	265	41	5
	3.9	1.09	244	39	4
11. judge's charge to jury	4.1	0.94	64	233	14
	4.1	0.96	60	213	14

Note 1: The first entry in a cell is from the Duplicated Data File, and the second entry is from the Unduplicated Data File.

Note 2: The mean, standard deviation (S.D.), and N statistics in this table are based on responses coded as follows: 5=Excellent, 4=Good, 3=Adequate, 2=Less than Adequate, 1=Poor. 'Not Applicable' responses are excluded from these statistics.

Table 1. Panel (b). *Lawyer's Questionnaire* item statistics.

Scale 2: Judicial Management Skills					
Item (paraphrased)	Mean	S.D.	N	'Not Applicable' (#)	missing (#)
1. moving proceeding, expeditious	4.2	0.95	296	10	5
	4.2	0.97	274	8	5
2. maintaining proper control	4.4	0.73	293	9	9
	4.4	0.75	270	8	9
3. doing necessary homework	4.1	1.05	287	14	10
	4.1	1.07	265	12	10
4. no unnecessary delay	4.2	0.96	300	8	3
	4.2	0.98	277	7	3
5. adequate time given	4.3	0.87	298	9	4
	4.3	0.88	276	7	4
6. resolving problems	4.0	1.12	267	38	6
	4.0	1.14	250	31	6
7. effecting compromise	3.9	1.12	167	134	10
	3.9	1.13	155	122	10
8. industriousness	4.2	0.96	246	54	11
	4.2	0.98	227	49	11

Table 1. Panel (c). *Lawyer's Questionnaire* item statistics.

Scale 3: Comportment					
Item (paraphrased)	Mean	S.D.	N	'Not Applicable' (#)	missing (#)
1. attentiveness	4.4	0.79	307	0	4
	4.4	0.80	283	0	4
2. courtesy	4.3	0.96	308	0	3
	4.3	0.98	284	0	3
3. compassion	4.1	1.05	250	54	7
	4.0	1.06	231	49	7
4. patience	4.1	1.07	299	8	4
	4.1	1.08	275	8	4
5. absence of arrogance	4.2	1.02	304	4	3
	4.2	1.00	281	3	3
6. absence of bias	4.4	0.83	290	15	6
	4.5	0.84	267	14	6
7. even-handed... litigants	4.3	0.98	283	24	4
	4.3	1.00	261	22	4
8. even-handed... attorneys	4.3	0.96	305	1	5
	4.3	0.98	282	1	4

(b) Questionnaire Scale Statistics

A series of analyses were conducted on the aggregate mean scores for each of the questionnaire's scales (or sections): Legal Ability, Judicial Management Skills, and Comportment. For these analyses, item responses were coded as 5=Excellent, 4=Good, 3=Adequate, 2=Less than Adequate, 1=Poor. 'Not Applicable' responses were excluded from the computations of scale means (i.e., treated the same as blanks or missing data). Table 2 summarizes descriptive statistics for the three scales.

Table 2. Descriptive statistics for *Lawyer's Questionnaire* scales

	Mean	Median	S.D.	N	Proportion 'NA'	Proportion 'NA' or missing
Legal Ability	4.0	4.1	0.93	305	19%	21%
	4.0	4.1	0.95	281	19%	21%
Judicial Management	4.2	4.3	0.84	307	11%	13%
	4.2	4.4	0.85	284	11%	13%
Comportment	4.3	4.6	0.84	309	4%	6%
	4.3	4.6	0.85	285	4%	6%

Note 1: The first entry in a cell is from the Duplicated Data File, and the second entry is from the Unduplicated Data File.

Note 2: The column "Proportion 'NA'" is the proportion of item responses of all possible responses (e.g., for Legal Ability, 11 x 311 is the number of possible responses in the Duplicated Data File) that were marked 'Not Applicable.' The column to the extreme right tracks the combined proportion of item responses that were 'NA' or missing (blank).

The occurrence of non-response (blanks) was generally small, accounting for 2% of all possible responses in each of the three scales (note the differences between the last two columns in Table 2). The occurrence of 'Not Applicable' responses, though, was substantial, especially for Scale 1 (Legal Ability) at 19% and Scale 2 (Judicial Management) at 11%. This implies, essentially, that aggregate scores for Scale 1 (whether one uses totals or means across the items as a scale score does not matter here) are being based on responses to just 79% (100% - 21%) of the items.

An extreme example might clarify the nature of the problem that large amounts of 'Not Applicable' responses or missing data can pose. For Scale 1 which is comprised of eleven items, a respondent could mark 'NA' to ten of the items and give a "4" rating to the remaining item. The mean score for Scale 1 from that respondent would be 4.0, even though based on just one item.

Inter-correlations among the three scales are shown in Table 3. Generally, one expects that the scales from the same questionnaire will be moderately inter-correlated. Relatively "low" inter-correlations, say roughly .40 or lower, may indicate that the scales (and the underlying content) do not really "go together." Relatively "high" correlations, say .90 or higher, may indicate that the scales are measuring much the same thing, and may not actually be psychometrically distinct scales. The correlations in Table 3 could be characterized as of "moderate" magnitude. The higher correlation between Legal Ability and Judicial Management (.87), as compared with those for Legal Ability and Comportment (.63) or Judicial Management and Comportment (.73), seems in line with what one should expect considering the nature of the content "tapped" by the three scales.

Table 3. Intercorrelations of *Lawyer's Questionnaire* scales.

	Legal Ability	Judicial Management	Comportment
Legal Ability	1.00	0.87 (303)	0.63 (305)
Judicial Management	--	1.00	0.73 (307)
Comportment	--	--	1.00

Note: The correlations shown are Pearson product-moment correlation coefficients. The data analyzed were from the Duplicated Data File. Numbers in parentheses are the N for which data on pairs of scales were available for the computation. (Nearly identical correlation coefficients, within one digit in the hundredths decimal place, were obtained from the Unduplicated Data File.)

An attempt was made to explore the scale structure of the *Lawyer's Questionnaire* using a variable clustering procedure (SAS Varclus) which is similar in purpose to factor analysis. However, clustering and factor analysis procedures, generally, omit records containing *any* missing item data. (Recall that for the item-based analyses, 'Not Applicable' responses were recoded as missing values.) An insufficient number of records without any missing item data (only 39 in the Duplicated Data File) were available, precluding meaningful analysis.

Estimates of scale reliability are extremely important. Unreliable measures are worthless. On the other hand, reliability does not guarantee validity. Reliability is a necessary but not sufficient condition for validity. Table 4 presents values of Cronbach's alpha coefficient, an estimate of scale reliability. Values of 1.00 are maximum and almost never occur except in simulation studies with artificial data. The reliability estimates found here are very high and could be characterized as "excellent."

Table 4. Reliability (internal consistency) of *Lawyer's Questionnaire* scales.

	Alpha Coefficient	SEM
Legal Ability	0.97	0.17
Judicial Management	0.94	0.21
Comportment	0.96	0.16

Note: The values shown are from analysis of the Duplicated Data File. 'SEM' is the standard error of measurement which is useful for considering the accuracy or precision of an individual's score. (Analysis of the Unduplicated Data File yielded identical results for the alpha coefficients except for Judicial Management, 0.93.)

Similar to variation in estimates that occur due to sampling, error is inherent in measurement processes and results in some (random) variation from one measurement to another. The standard error of measurement (SEM) can be used to index the amount of measurement error associated with an *individual* score. Roughly speaking, one might say that an individual true score or rating is likely to be within plus or minus two SEM's (with about 95% confidence). [Technically this is incorrect, but for practical purposes, the exposition here is "accurate."] For the Legal Ability scale, for example, a given lawyer-respondent's rating of, say, 4.0 would likely vary between about 3.7 and 4.3 ($4.0 \pm 2 \times 0.17$) if we could somehow repeat the rating process many times with no memory effects, fatigue effects, etc.

However, the precision of individual attorney-respondent ratings is not at issue here. Rather, for judicial profiles, minimum aggregates of 20 individual scores or ratings have been specified. How reliable are the aggregated scores? Is an aggregated *mean* score of 4.2 "really" different than a 4.3? As it turns out, the answer is probably not. Using a formula developed by Feldt and Brennan (1989, p. 127, formula #60), we used the available data to estimate the reliability of the aggregated mean scores, and then, we also computed an SEM for the aggregated mean scores. The following table shows the results obtained.

Table 5. Reliability of mean (average) scores from *Lawyer's Questionnaire*.

	Reliability of Mean Scores	SEM of Mean Scores
Legal Ability	0.79 (0.80)	0.16 (0.18)
Judicial Management	0.81 (0.83)	0.15 (0.16)
Comportment	0.87 (0.87)	0.14 (0.15)

Note: The first values shown are from analysis of the Duplicated Data File. Values in parentheses were derived from the Unduplicated Data File. ('SEM' is an acronym for Standard Error of Measurement.)

The reliability estimates for mean scores in Table 5 take into account sampling variation, i.e., variation that arises from the particular group of raters (lawyer-respondents) who provided scores for each judge. The set of lawyer-respondents for each judge can be regarded as a sample, within a longitudinal perspective, of possible raters. Interpretations of the mean scores must consider such sampling-related variation or error.

Further, considering the very high reliabilities of the individual ratings, the "error" in judges' aggregate mean scores is mostly a function of sampling variation rather than measurement error. As a computational check, we computed the standard error of the mean (using traditional statistical methods) which accounts for sampling variability only. The values obtained were 0.01 to 0.02 larger than the SEM values in the table above; the difference probably arising from an acknowledged "small positive bias" in Feldt and Brennan's reliability formula (p. 127).

Assuming a relatively high degree of confidence is desirable for placing a confidence interval around a given mean score (say, 95% confidence), we would again use the criteria of plus or minus two SEM's. Based on the current Circuit Court data, the confidence interval would be found by taking the mean value ± 0.3 . Thus, one must be guard against attributing excessive precision to the mean ratings. Were it somehow possible to acquire many sets of 20 or more ratings for a given judge, we would likely find, for instance, that a mean for Legal Ability of 4.0 obtained from one of the samples, would range from 3.7 to 4.3 in the other samples.

A related but somewhat different statistical issue is that of statistically significant mean differences. The questions might arise "Is Judge #1's mean at Time 1 significantly different from his/her mean at Time 2?" or "Is Judge #1's mean significantly different from Judge #2's?" Based on variance estimates from the current Circuit Court data, and assuming that samples of $n = 20$ respondents were involved, we estimated the size of the difference in means (per scale) required for statistical significance at the $\alpha = 0.05$ rejection level. (Traditional statistical formula for testing means from independent samples were used.) For Legal Ability, Judicial Management Skills, and Comportment, the differences obtained were 0.6, 0.5, and 0.5, respectively. *As a rough rule-of-thumb, then, means would have to be different by at least 0.5 of a scale point (e.g., half the distance between "Good" and "Excellent" on the rating scale) before one might conclude that a statistically significant mean difference had been found.* Said another way, mean differences smaller than 0.5 may be due simply to "luck of the draw" in terms of which 20 attorney-respondents happened to be selected into the samples. Sample size plays the major role here. In concept, one could reliably detect smaller mean differences by using larger samples; but in practice that may not be feasible.

(c) Results by Judge

Mean values on each of the three scales for each of the 12 Circuit Court judges were examined. In brief summary, the two primary findings were as follows: (1) One-third of the judges lacked the required number of questionnaires (i.e., minimum N of 20) that Planning and Evaluation Division staff would need in order to develop a judicial profile; and (2) using duplicated or unduplicated data, per individual judge, made little substantive difference. Means were identical for eight of the 12 judges when using either data set, and differed by no more than 0.2 for the other four judges. In addition, where differences occurred, there was no indication, overall, of more or less favorable results from either data file.

(d) Exploration of Relationships

(d.1) 'Not Applicable' Responses and Type of Proceeding

Previously we noted that some items received markedly high rates of 'Not Applicable' (NA) responses. This can be an indication that such an item's content may not be relevant, generally, or specifically, to the scale to which it is assigned or that it does not function as intended in the overall measurement effort. For purposes of the present exploratory analysis, we focused on those items marked NA by 10% or more of the respondents.

Ten items were marked NA by 10% or more respondents. Item "11," "Judge's charge to the jury," an extreme example, was marked NA by 75% of the lawyer-respondents. We suspected that the occurrence of NA responses, and the relevance of questionnaire item content, might be related to the type of the proceeding. Other analysis of the Duplicated Data file (reported in *Appendix B*) found that 29% of the proceedings were trials and 71% were motions.

The fourth column in Table 6, labelled "% 'NA' w. Proceeding a Motion," refers to data extracted from a series of cross-tabulations (not shown here) wherein a given item (say item "3" from the Legal Ability scale) was cross-tabulated with Q4 of "Section 4: Background Characteristics." Q4 or Question "4" asks the respondent to describe the outcome of the proceeding. The figure in the fourth column is the number of NA responses to the item (listed in column "2") *if the proceeding was a motion*, divided by the total number of NA responses to the item (expressed as a per cent). Since the "base rate" of motions is 71%, one would expect the ratios in column "4" to be close to that value, *if there were no relationship between giving an NA response and the type of proceeding*.

Table 6. Items with relatively high 'Not Applicable' response rates related to motion proceedings

Scale	Item (paraphrased)	% 'NA' Response	% 'NA' w. Proceeding a Motion	More than expected?
Legal Ability	3. know rules of evidence	32%	92%	yes
	6. giving reasons for rulings	10%	87%	yes
	8. adequacy of findings	41%	72%	no
	9. clarity of decision	13%	62%	no
	10. completeness of decision	13%	66%	no
	11. judge's charge to jury	75%	86%	yes

Scale	Item (paraphrased)	% 'NA' Response	% 'NA' w. Proceeding a Motion	More than expected?
Judicial Management	6. resolving problems	12%	82%	yes
	7. effecting compromise	43%	85%	yes
	8. industriousness	17%	91%	yes
Comportment	3. compassion	17%	91%	yes

Note: Data for this table are from the Duplicated Data File.

Results for seven of the ten items flagged out for this analysis indicate that the frequency of occurrence of NA responses to those seven items is associated with the type of proceeding, i.e., motions. A series of follow-up chi square tests confirmed what seems intuitively evident by inspection: the seven items marked with "yes" in column "5" each had statistically significant chi square values using a rejection level of $\alpha = .05$; and, the three items with "no" in column "5" did not. *These seven items (with "yes" in column "5") should be carefully reviewed in terms of their content relevance to evaluating judges when the proceeding is a motion.*

A parallel analysis using the Unduplicated Data File obtained the same substantive results, with one exception. That exception occurred for Item 6 of the Judicial Management scale ("Resourcefulness and common sense in resolving problems arising from the proceeding"), which did not have a statistically significant chi square (using either the standard chi square procedure or Fisher's exact test) at the $\alpha = .05$ rejection level.

(d.2) 'NA' Responses, Missing Data, and Scale Means

A possible concern is that the occurrence of NA responses and omitted items (blanks), which are presumably treated as missing data values when computing judges' means for each of the three scales (as was done herein), may bias the aggregated mean values. (A related concern, addressed elsewhere below, is more conceptual in nature--if a respondent marks either NA or skips 50% or more of the items in a scale, should the data from that respondent for that scale be used at all?) We explored the former issue by examining the correlation coefficients between the scale means for each scale and the corresponding (a) number of NA's marked, and (b) total number of items marked NA or omitted. Data from the Duplicated Data File were used for the analysis. One would hope that higher (or lower) mean scores were unrelated to the frequency of occurrence of NA's or omitted items.

Five of the six correlations generated were near zero in magnitude and not statistically significant (using a rejection level of $\alpha = .05$). The correlation coefficient between the mean values for Scale 2 (Judicial Management Skills) and the combined total number of items marked NA or omitted in that scale was -0.11 ($n = 307$, statistically significant for $\alpha = .05$). This finding implies that there is a *slight* tendency for lower Judicial Management Skills scores to be associated with higher frequencies of items marked NA or omitted.

Substantively identical findings were obtained from a parallel analysis of the Unduplicated Data File. All correlations were near zero, except, as above, for that between the mean values for Scale 2 (Judicial Management Skills) and the combined total number of items marked NA or omitted: -0.14 ($n = 284$, statistically significant for $\alpha = .05$).

Practically, however, the effect of this relationship is likely trivial and ignorable, given the relatively small size of the correlation coefficient--unless a "pile up" of many records with many NA and omitted items occurs for a specific judge. *Generally, then, it does not appear that the occurrence of NA responses and omitted items (blanks) is associated with any systematic increase or decrease in the ratings.*

Another exploratory analysis looked further at the potential issue of bias in judges' mean scores due to the occurrence of NA responses and omitted items. We examined the effects of deleting those questionnaire records with relatively large amounts of NA responses and omitted items. The deletion criteria used (with the Unduplicated Data File) was: If 'Not Applicable' responses or missing data (blanks) comprised 50% or more of a given scale from a given attorney-respondent, all data for that scale was deleted from that record. In effect, unless an attorney-respondent marked responses corresponding to "1," "2," "3," "4," or "5" for at least 50% of the items in a scale, the data from that respondent for that scale were ignored. Findings indicated clearly that *the effect of attempting to further control for NA and missing data is slight* -- 0.1 mean rating point on a scale of "1" to "5" for a given judge in comparison with the Unduplicated Data File. *The occurrence of NA responses and omitted items (blanks) does not appear to be associated with any notable increase or decrease in mean ratings for individual judges.*

(d.3) Selected Variables and Scale Means

In this final exploratory section we examine relationships between scale means and three variables that the Judicial Performance Program could directly control (if needed) via its selection of respondents and/or records for profile analysis: (a) the attorney-respondent type, (b) the type of proceeding (civil or criminal), and (c) outcome of the proceeding. Data from the Duplicated Data File were used for the following analyses.

Table 7. Variation in means by attorney-respondent type.

	Attorney-Respondent Type					
	Prosecutor/ Plaintiff			Defense		
	Mean	S.D.	N	Mean	S.D.	N
Legal Ability	3.9	0.96	142	4.1	0.91	163
Judicial Management	4.1	0.87	144	4.2	0.80	163
Comportment	4.2	0.83	144	4.3	0.85	165

On each of the three scales, there seems to be a tendency for defense attorneys to rate judges slightly higher than prosecutors' or plaintiffs' attorneys (Table 7). However, the data are subject to sampling variation. A series of *t*-tests for differences between the means found *no statistically significant differences* (at a rejection level of $\alpha = .05$). Substantively identical conclusions were obtained from parallel analysis of the Unduplicated Data File.

Table 8. Variation in means by type of proceeding.

	Type of Proceeding					
	Criminal			Civil		
	Mean	S.D.	N	Mean	S.D.	N
Legal Ability	4.2	0.82	128	3.9	0.99	177
Judicial Management	4.2	0.78	129	4.2	0.88	178
Comportment	4.1	0.99	129	4.4	0.69	180

The results for type of proceeding are mixed (Table 8). Attorneys in criminal proceedings appear to rate judges higher on Legal Ability than do their counterparts in civil proceedings.

The difference (4.2 v. 3.9) is statistically significant at the .05 level (using a *t*-test procedure for unpaired samples with unequal variances). No difference occurred for Judicial Management. For the Comportment scale, the direction of the difference (4.1 v. 4.4) reversed, with attorneys in civil proceedings providing the higher mean rating. The difference in means for Comportment is also statistically significant at the .05 level. (As in previous analyses, substantively identical conclusions were obtained from parallel analysis of the Unduplicated Data File.)

Table 9. Variation in means by outcome of proceeding.

	Outcome of Proceeding					
	Trial Won			Trial Lost		
	Mean	S.D.	N	Mean	S.D.	N
Legal Ability	4.2	0.62	31	3.7	0.98	22
Judicial Management	4.3	0.59	31	4.0	0.96	21
Comportment	4.4	0.65	31	4.0	1.09	22
	Motion Granted			Motion Denied		
	Mean	S.D.	N	Mean	S.D.	N
Legal Ability	4.2	0.87	114	3.9	1.04	71
Judicial Management	4.3	0.77	115	4.1	0.91	72
Comportment	4.3	0.79	116	4.1	0.94	72

Note: The outcomes shown in the table above were intentionally selected for the dichotomous contrast provided. The outcomes shown represent about 79% of the data available for this item (question "4" of "Section 4: Background Characteristics"). Other outcomes, not included in the table, were: Trial Won in Part, Trial Dismissed, Trial Other, and Motion Other.

It may seem clear that attorneys with successful trials (won) tended, on average, to rate judges slightly higher than did their less successful counterparts (Table 9). However, the data are subject to sampling variation. The results of *t*-tests indicate that *for trials, only the difference in means for Legal Ability* (4.2 v. 3.7) *was statistically significant* ($\alpha = .05$). *For motions, only the mean difference for Judicial Management was statistically significant* at the .05 level (but differences for Legal Ability and Comportment were "borderline," i.e., very close to reaching the .05 level). [The latter results, for motions, are not intuitively obvious and cannot be obtained simply by examining the differences between means in Table 9. Statistical significance also depends upon the *n* size and the standard deviations. Thus, for instance, had somewhat larger samples been available, all the mean differences could have been statistically significant.]

Results of the parallel analysis with the Unduplicated Data File were similar in all respects to those just described except one: the mean difference for Legal Ability, for proceedings that were motions, was also statistically significant.

The findings of this section are useful in the following way. If Planning and Evaluation Division staff select or reject survey records for inclusion in profiles (e.g., in order to eliminate "duplicates"), that ought to be done with the knowledge that the type of proceeding (civil or criminal), and outcome of the proceeding matter in terms of mean scores. If selection is necessary and if choices are available, selection should attempt to maintain the representativeness of type and outcomes of the proceedings typical of *each* judge.

(e) Partial Confirmation with District Court Data

Upon our request a series of small "in progress" data files were provided by the Planning and Evaluation Division for use in the current evaluation. Files for several courts and judicial circuits were provided, but only two of the data files contained more than 30 questionnaire records, and any analysis of such small data files would have been extremely speculative. Although probably best viewed as "preliminary" data, we did have available moderate amounts of questionnaire data from the District Court of the First Circuit: 116 records in a "duplicated" data file (je5.d, dated 12/30/94) and 91 records in a corresponding unduplicated version of the larger file (je5.e, dated 1/11/95). Analysis was conducted using the larger "duplicated" data file ($n = 116$).

Here, for the sake of brevity, we will report only two findings that are central to our studies of instrument validity reported in the preceding sub-sections.

For the District Court of the First Circuit data, internal consistency reliability coefficients of the *Lawyer's Questionnaire*, as indexed by Cronbach's alpha, were .99, .96, and .97, for Legal Ability, Judicial Management Skills, and Comportment, respectively. These reliability estimates are very similar to the reliability estimates (Table 4) derived from the Circuit Court of the First Circuit data. Thus, reliability findings from both the Circuit Court and the District Court of the First Circuit consistently indicate that, from a measurement perspective, the scores or ratings provided by the *Lawyer's Questionnaire* are highly reliable.

The other issue we examined in some detail was that of high levels of 'Not Applicable' (NA) responses to some items. (The occurrence of relatively large numbers of respondents marking 'Not Applicable' to certain items *may* be an indication of content validity problems with those items.) Table 10 below summarizes what was found, listing those items marked NA by 10% or more of the District Court lawyer-respondents. For comparison, the corresponding results (from Table 6) for Circuit Court ("duplicated" data file) are also shown.

Table 10. Items with relatively high 'Not Applicable' response rates in District Court and in Circuit Court

Scale	Item (paraphrased)	Percent Marked 'NA'	
		District Court	Circuit Court
Legal Ability	3. know rules of evidence	--	32%
	6. giving reasons for rulings	--	10%
	8. adequacy of findings	16%	41%
	9. clarity of decision	--	13%
	10. completeness of decision	--	13%
	11. judge's charge to jury	95%	75%
Judicial Management	3. doing necessary homework	39%	--
	6. resolving problems	11%	12%
	7. effecting compromise	37%	43%
	8. industriousness	11%	17%
Comportment	3. compassion	--	17%
	8. evenhanded w. attorneys	13%	--

Notes. "--" designates rates of 'NA' response under 10%.

Both the District and Circuit Courts included here are of the First Judicial Circuit only.

Our admittedly *ex post facto* explanation of these findings hinges on the (plausible) assumptions that a lawyer-respondent's ability to provide a rating other than 'Not Applicable' to a given item is related to (1) the nature of the proceeding and, thereby, (2) the opportunity the proceeding affords the respondent to observe or experience first-hand the judicial behavior described by the item. Generally, then, one would expect trials, more so than motions or arraignments, to provide a more comprehensive opportunity for judicial evaluation by lawyer-respondents via the *Lawyer's Questionnaire*.

Previously in the present report we noted, for the Circuit Court questionnaire data, that 29% of the proceedings were trials and 71% were motions.

For the District Court data we found that of the 110 respondents who indicated the type of proceeding (six did not provide a response to the item asking about the outcome of the proceeding, which we used to categorize the kind of proceeding), all or 100% of the proceedings were trials! Thus, for the data shown in the immediately preceding table, it appears that 100% of the District Court proceedings were trials, whereas only 29% of the Circuit Court proceedings were trials. Thus, the District Court data are generally consistent

with our previous analysis with Circuit Court data which found that the frequency of occurrence of NA responses to be associated with the type of proceeding, with proceedings that were motions resulting in higher rates of NA response.

(2) *Results of Content Analysis*

The five-page *Lawyer's Questionnaire* consists of five sections, including one for background information on the respondent and another for comments. The remaining three sections constitute the substantive "dimensions" of the questionnaire. These include Legal Ability, comprised of 11 items, Judicial Management Skills, comprised of 8 items, and Comportment, also 8 items. The scale is a 5-point Likert-type continuum used across all three dimensions. The anchor descriptors for the five categories are: "Excellent," "Good," "Adequate," "Less Than Adequate," and "Poor." A sixth category, "Not Applicable," was also used.

The Hawaii *Lawyer's Questionnaire* is an adaptation of a similar instrument with the same name from New Jersey's Judicial Performance Program. The same number of categories are used with one substitution in the anchor descriptor -- "More Than Adequate" is replaced by "Good". There are some differences in the number of items per category (generally fewer). Thus, the total number of items rated in the Hawaii *Lawyer's Questionnaire* totals 27 in comparison to New Jersey's 36. Some noteworthy changes in the adaptation, outside of the elimination of nine items, are the transfer of the item "Industriousness" from Comportment to Judicial Management Skills, and a new item in Comportment, "Compassion."

Table 11 on the following pages provides a comparative summary of instrument characteristics for judicial performance evaluation questionnaires used in Hawaii, New Jersey, and Connecticut.

Hawaii's goals, governing entities, evaluation methods, and system of judicial retention are similar to these other two states. Both states, moreover, have been pioneering the planning and development of judicial performance programs since the early 1980's, and were among the few that first participated in the National Center for State Courts judicial performance evaluation project.

An inspection of Table 11 reveals that Hawaii limits use to a single instrument whereas the other two states have experimented with the use of different instruments for different court environments and for different evaluation information sources. New Jersey, for example, utilizes two types of lawyer questionnaire, one specifically geared for use in "high volume and special proceedings" courts. Connecticut is, at this time, in the midst of piloting a similar "high volume" type instrument. Both New Jersey and Connecticut at one time were including jurors as an additional evaluation information source. Although Connecticut continues to do so, New Jersey has dropped the use of the juror questionnaire. Furthermore, Connecticut's use of appellate and administrative judges as additional sources of evaluation information has been discontinued due to the "...limited information appellate and administrative judges can provide about courtroom conduct of superior court judges (State of Connecticut, 1991, p.2). All of the instruments have a section for soliciting comments; all of the instruments except for Connecticut's *Juror's Questionnaires* collect information on background characteristics.

Table 11. Comparative summary of instrument characteristics for judicial performance evaluation questionnaires used in selected state programs.

Program	Instrument	Category	Items	Backed?	Comment?	Scale Characteristics
Hawaii Judicial Performance Program	Lawyer's Questionnaire	Legal ability	11	Yes	Yes	5-point Likert
		Judicial mgt. skills	8			5-point Likert
		Comportment	8			5-point Likert
		Total	27			
New Jersey Judicial Performance Program	Lawyer's Questionnaire (major adversarial)	Legal ability	11	Yes	Yes	5-point Likert
		Judicial mgt. skills	13			5-point Likert
		Comportment	12			5-point Likert
		Total	36			
	Lawyer's Questionnaire (high volume)	Legal ability	7	Yes	Yes	5-point Likert
		Judicial mgt. skills	13			5-point Likert
		Comportment	12			5-point Likert
		Total	32			
	Juror's Questionnaire	Untitled; mix of comportment & mgt. skills	9	Yes	Yes	Not Available
	Appellate Judge's Questionnaire	Legal ability	13	Yes (1 only)	Yes	5-point Likert
		Judicial mgt. skills	3			5-point Likert
		Comportment	5			5-point Likert
		Total	21			

Program	Instrument	Category	Items	Backlog?	Comment?	Scale Characteristics
Connecticut Judicial Performance Evaluation Program	Attorney's Questionnaire (1991)	Legal ability	?	Yes	Yes	4-point Likert-type
		Judicial mgt. skills	?			4-point, 5-point Likert-type
		Demeanor	?			4-point, 5-point Likert-type
		Total	36			
	Attorney's Questionnaire (1990)	Legal ability	10	Yes	Yes	3-point continuum
		Judicial mgt. skills	13			3-point continuum
		Demeanor	7			3-point continuum
		Total	30			
	Juror's Questionnaire (1991)	Attitude	7	No	Yes	5-point Likert
		Untitled; mix of demeanor & mgt. skills	7			4-point Likert-type
		Total	14			
		Untitled; mix of demeanor & mgt. skills	10			3-point continuum
	Juror's Questionnaire (1984)					

Connecticut serves as a useful contrast in the area of scale characteristics. Earlier versions of both the *Attorney* and *Juror Questionnaires* were dropped in favor of revised questionnaires that have larger ranges in the scale continuum. The initial use of a 3-point scale consistently throughout the questionnaires was replaced by mixed 4- and 5-point Likert-type scales. In contacting Connecticut's Judicial Evaluation Administrator, we confirmed earlier hunches that the anchor descriptors for the early versions of the questionnaires were not adequate for purposes of judicial performance review. Also noteworthy was the abandonment of the intact section organization by construct (e.g., Legal Ability). Items in the revised questionnaires are organized by response type categories instead.

The juror questionnaires for New Jersey and Connecticut tend to tap areas very similar to those found in their respective lawyer-type questionnaires. Most if not all the items ask for ratings related to comportment and courtroom management skills. Only the single item "charge to the jury" is one that jurors are asked to rate in the area of legal ability.

One notable difference is the absence of court and case identifying information in both Hawaii and New Jersey's questionnaires. Hawaii, for example, uses a separate "Judicial Survey Information Form" to keep tabs on such information. The *Questionnaire* itself does not contain information about the judge's identity, case number, and so forth. This procedure, requiring two forms, and keeping separate court and case identifying information from respondent ratings is preferable, because it maximizes the protection of confidentiality on the part of judges, and does not unnecessarily jeopardize the validity of ratings (due to respondents being not entirely convinced that confidentiality measures are adequate).

Two other findings of the content analysis should be mentioned here. The first is a relatively minor oversight, but should be corrected. Item 2 in the "Background Characteristics" section of Hawaii's *Lawyer's Questionnaire* includes in the response categories, overlapping numbers in the range of years respondent "...practiced law." To avoid ambiguity, the range in the response categories should not overlap. The second is an observation that Hawaii's Lawyer's Questionnaire, in contrast to those reviewed for other states, has been revised and formatted in Optical Mark Reader (OMR) scannable form. In this regard the instrument has improved both in aesthetic form and in function.

(3) Results of Interviews

As described previously in the "Methods" section of the present report, a series of 16 telephone interviews were conducted by the evaluators with ten (10) administrative judges and senior judges, all of whom were participants in the Judicial Performance Program, and eight (8) members of the Special Committee on Judicial Performance. Two of the Special Committee members were also among the ten judges in the first group of interviewees. Consequently, the total number of individual interviewees was 16.

The interviews provided an excellent opportunity to obtain information directly from the program's participants (judges) as well as the program's developers/planners (Special

Committee members). The interview questions focused on issues of instrument validity and the evaluation process.

The summary of findings that follow were obtained from analysis of interview notes, which primarily involved category coding of responses and tabulation. A complete set of analyses for each interview question (and follow up questions, where applicable) are provided in **Appendix D**. Below we present essential highlights that are most directly relevant to issues of validity.

Two key questions, which could reveal "fatal flaws" in program procedures, asked "Does the Judicial Performance Program's implementation interfere with normal courtroom practices?" and "Does the Judicial Performance Program infringe upon the independence or integrity of the judiciary?" (Unlike all other questions used in the two interview protocols--see **Appendix C** for sample copies of the protocols--these two questions were asked only of judges or attorneys.)

- Of the 14 judges and attorneys responding to the first question, 12 said "No," that it did not interfere with normal courtroom practices. One interviewee, a judge, said "Yes," and went on to say that the program "Makes us perform better!" Another interviewee said, due to lack of familiarity with the program, "Don't know."
- Of the same 14 judges and attorneys responding to the second question, 12 said "No," that the program did not infringe upon the independence or integrity of the judiciary. Two respondents gave a "Don't know" type of response.

The interview findings seem to clearly indicate that the Judicial Performance Program does not negatively interfere with normal courtroom practices, nor is it perceived currently as infringing upon the independence or integrity of the judiciary. It might be noted, though, that several interviewees qualified their "No" response to the infringement question. Three mentions, for instance, were made that public release or media use could result in infringement.

Another question asked interviewees (n = 16) about the appropriateness, for Hawaii, of five purposes for which judicial evaluation programs similar to Hawaii's have been used. The number of interviewees endorsing "appropriate" or "inappropriate" or "other" for each of the five purposes were as follows:

Table 12. Appropriate purposes of the Judicial Performance Program?

Purpose	Appropriate	Inappropriate	Other
self-improvement	16	0	0
designing judicial education programs	16	0	0
assignment of judges	13	2	1
promotion decisions	10	5	1
retention decisions	15	0	1

Again, it should be noted that some interviewees qualified the responses shown in Table 12. *Appendix D* contains additional details, but two of the qualifications should be mentioned here. First, several interviewees expressed reservations regarding the use of Judicial Performance Program information for retention decisions (e.g., preserving confidentiality). Second, "promotion decisions" may not really be a relevant purpose in Hawaii. If we understood correctly what several interviewees told us, it seems that the judiciary in Hawaii has an appointment process, but no promotion process *per se*.

Interestingly, *the 16 interviewees endorsed as appropriate both of the "low stakes" uses, self-improvement and judicial education, without any qualifying comments*. The last three purposes, assignment, promotion, and retention are "high stakes" because of direct impact on career, and each were accompanied by related qualifying statements or reservations from some of the interviewees. Thus, *while all purposes, with the exception of promotion, were largely endorsed as "appropriate," the high stakes purposes were sometimes given qualified endorsements*.

Related to content validity, one of the interview questions asked respondents ($n = 16$) to estimate what proportion (in percent) of a judge's responsibilities are covered by the items in the *Lawyer's Questionnaire*. This is a critically important question. Put simply, if the *Lawyer's Questionnaire* is to be used to make inferences about the "job" performance of a judge, then the items must relate to the duties and responsibilities of the judge's assignment. And, if important duties and responsibilities are not covered by the questionnaire, then corresponding limitations must be attached to interpretations and uses made of the resulting information.

- Of the 13 interviewees who were able to supply a quantitative response to the question, all responses were in the range of 75% to 100%, and ten (10) of these 13 interviewee's responses corresponded to a coverage rate of at least 90%.

While it appears that most interviewees viewed the Lawyer's Questionnaire as covering almost completely the responsibilities of a judge, there is some question about which kinds of proceedings, and judges, the instrument is appropriate. One interviewee noted that the questionnaire may not have appropriate content for evaluating administrative and motions judges.

The last item in the interview was open-ended and asked for any comments about the program's strengths and weaknesses. A complete listing of all comments provided is given in *Appendix D* (pages D6-D8). Below, in summary form, are some selected highlights:

Strengths

- The program will help judges by providing information relevant to self-improvement.
- The program provides balanced feedback, not just idiosyncratic complaints.
- The program has been carefully, incrementally developed, has high quality staff, and has solid leadership support.

Weaknesses/Concerns

- Some attorneys are still concerned about the adequacy of the program's confidentiality provisions.
- Court staff's listing of the attorneys to be surveyed could conceivably bias the selection of the evaluator-respondents selected.
- The fit of the *Lawyer's Questionnaire* and program procedures with district court proceedings may be poor.

B. Validity of the Evaluation Process

(1) *Results of Personnel Evaluation Standards Comparisons*

Findings from the application of the 21 Personnel Evaluation Standards are summarized in Table 13. Final ratings of the degree to which the Judicial Performance Program's procedures do or do not meet the Standards are included, as well as annotations on the strengths and/or weaknesses observed. Where appropriate, recommendations specific to the improvement of the individual standard are included also. Five of the 21 Standards were not rated because the Program has not as yet implemented the functions tapped by those Standards. These functions are related largely to the use and dissemination of evaluation results.

Frequency tabulations of ratings, by domain and the degree to which Standards were met, are shown in Table 14.

Table 14. Judicial Performance Program ratings, by domains, using the Personnel Evaluation Standards adapted for the judiciary.

Domain	# Items	Personnel Evaluation Standards				
		Ratings				
		Met	Part Met	Not Met	?	Mean Rating
Propriety	5	2	2	0	1	2.5
Utility	5	1	1	0	3	2.5
Feasibility	3	2	0	0	1	3.0
Accuracy	8	6	2	0	0	2.75
Total	21 (100%)	11 (52%)	5 (24%)	0 (0%)	5 (24%)	2.7

Despite the omission of five Standards for which ratings were not attempted, more than half of the 21 Standards were rated as "Met" (11 of 21; 52%). Another five (24%) were "Partially Met," and none were rated "Not Met." Most (three-fifths) of the omitted Standards belong to the Utility domain, and is directly related to the deferring of evaluation reports.

Table 13. Summary of findings and recommendations, adapted Personnel Evaluation Standards

Standard	Strengths	Weaknesses	Met=3 Part Met=2 Not Met=1	Recommendations
P-1: Service Orientation	Emphasis on improvement of judges, individually & of the judiciary overall; covers major functions, including legal competence, judicial management skills & comportment.	Multiple goals & purposes of performance evaluation may not be entirely compatible.	3	Review priorities, identified goals & purposes, and how multiple purposes may impact the primary goal of improving judicial performance.
P-2 Formal Evaluation Guidelines	Court rule established; drafted policies & procedures manual useful for formal review & training.	Policies & procedures manual needs updating and reorganizing for optimal use.	2	Reorganize & update manual.
P-3 Conflict of Interest	Policies & procedures formulated to protect & uphold independence, integrity of judiciary's function	Selection of cases not clearly defined operationally; can introduce too much discretion on part of court staff.	2	Use objective criteria in specifying how cases are to be selected; or transfer function to other support staff.
P-4 Access to Personnel Evaluation Reports	Protective measures to limit access to individuals with legitimate needs seem adequate.		3	
P-5 Interactions with Evaluatees	Emphasis on self-improvement rather than discipline; some form of follow-up individually anticipated (e.g., debriefings or consultations).		?	Not yet implemented

Standard	Strengths	Weaknesses	Met=3 Part Met=2 Not Met=1	Recommendations
U-1 Constructive Orientation	Intent of Program's goals & purposes constructive.	As with Standard P-1, multiple goals & purposes potentially at odds, competing.	?	Emphasize purposes of improving judicial performance & continuing education; allow Program to mature before expanding to other uses.
U-2 Defined Uses	Intended uses identified & described adequately in Program goals & Court Rule 19.	Program lacks overall long-term plan; unclear as to implementation timetable.	2	Outline & develop long-term strategic plan.
U-3 Evaluator Credibility	Lawyers considered single best source of evaluation information on courtroom performance; most credible source utilized.		3	
U-4 Functional Reporting	Sample "profiles" based on evaluation results for individual judges appear to be useful feedback; aggregate summaries based on all ratings also informative.	Interpretation of results may at times be difficult.	?	Not implemented yet; do consider supplementing computed summary statistics with graphic displays; explore approaches that allow automation.
U-5 Follow- Up and Impact		Unclear as to how evaluations will be followed up (e.g., debriefings with & recommendations by Chief Justice, administrative judges, established commission, etc.)	?	Not implemented yet; awaiting findings of evaluation(s).
F-1 Practical Procedures	Cost-effective use of survey research methods.		3	

Standard	Strengths	Weaknesses	Met=3 Part Met=2 Not Met=1	Recommendations
F-2 Political Viability	Involvement of stakeholder groups in both Study Committee & Special Committee; responsibility vested in highest court.		3	
F-3 Fiscal Viability	Program funding & administrative support provided.	Adequacy of resources? (e.g., dependence on court staff for case selections).	?	Investigate whether dependence on court staff for case selection unbiased; if serious problems found, additional resources may be required.
A-1 Defined Role	Role, qualifications & responsibilities of judges generally understood by legal community.	For purposes of Program, specific duties & responsibilities in jurisdictions other than circuit court do not seem sufficiently accounted for by evaluation instrument; but probably okay for low stakes uses such as self-improvement & education.	2	Consider relative successes other state judicial performance programs may have had in using different instruments and methods for different jurisdictions.
A-2 Work Environment	Evaluators (i.e., respondents) have first hand knowledge of courtroom context.		3	
A-3 Document-ation of Procedures	Procedures and controls seem to ensure that actual = planned.		3	

Standard	Strengths	Weaknesses	Met=3 Part Met=2 Not Met=1	Recommendations
A-4 Valid Measurement	Appraisal of performance requires professional judgement; taps most credible source of information: lawyers who have actually observed case proceedings in which presided by judge.	Only singular source of information available; could be stronger with cross-validation by other methods.	3	Investigate use of other approaches that could bolster current practice (e.g., self-assessments, review of case records & videotapes, trained court observers).
A-5 Reliable Measurement	Analysis of Lawyer's Questionnaire shows high reliability (internal consistency).		3	
A-6 Systematic Data Control	Good overall monitoring of data collection & procedures; confidentiality not compromised; scannable forms developed; data processing support adequate.		3	
A-7 Bias Control	Matched questionnaires distributed to both parties, plaintiff/prosecutor & defendant; temporary deferring of reports provided check on bias control before results were released.		2	
A-8 Monitoring Evaluation Systems	Program itself subjected to necessary periodic evaluation.		3	Include future monitoring at key intervals; integrate with long-term planning.

The mean (average) ratings to the extreme right of Table 14 were computed for each of the four domains and for the overall total. Based on a 3-point scale, Standards that were "Met" were assigned a value of "3." Standards "Partially Met" were assigned a "2," and Standards "Not Met" were scored "1." Thus the highest possible mean rating was a 3.0 and the lowest possible was 1.0. The mean ratings ranged from 2.5 for Propriety and Utility, to a high of 3.0 for Feasibility. The mean rating for the overall total was 2.7. All in all, these mean ratings, too, are quite favorable. They serve as useful, quick summary indices on the domains. It should be noted, however, that the mean ratings are based on relatively small numbers of items especially within the Feasibility domain, and because of omissions, the Utility domain as well.

Whether meeting half of the 21 Standards should be considered adequate or not is difficult to say. According to Stufflebeam, a good rule of thumb might well be that personnel evaluation systems meeting at least half of the 21 Standards is adequate, providing there are no "fatal flaws" inherent in those Standards not being met. [Response to this very issue posed to Stufflebeam, who served as chair of the Joint Committee's development effort, in a workshop on the Personnel Evaluation Standards held in Honolulu, Hawaii, January, 1995.] A prime example of such fatal flaws is that of not meeting Standard A-4, "Valid Measurement." Another fatal flaw, one could surmise, is one that does not meet Standard U-1, "Constructive Orientation." Thus, if the personnel evaluation system neither measures performance accurately nor is constructive or useful, the system is destined to fail.

Considering that a quarter of the Standards were not yet included in the ratings, and that none were rated "Not Met," the Program appears to have held up quite well against these general, exemplary requirements. However, an important caveat should be mentioned here again that these Standards were examined under the guiding assumption that the Program's first and foremost objective is that of improving individual judges' performances. For example, if any of the "higher-stakes" objectives such as retention is instituted on a more or less equal basis as improving individual judges' performance, it is unknown, at least questionable, whether both objectives will be sufficiently compatible and can co-exist in an environment where results are used for decisions on retention as well as for self-improvement purposes.

(2) Results of American Bar Association (ABA) Guidelines Comparisons

Results from applying the 47 checklist items adapted from the ABA Guidelines are shown on the following pages (Table 15). These results are laid out in a format somewhat different than that for the Standards. This checklist format simply documents whether the Program appears to have addressed the item or not (yes/no) and the degree to which the particular procedural requirement or performance criteria has been met. A "Findings/Comments" column also was included for use in elaborating findings where needed.

Frequency tabulations of ratings, by Guideline category and the degree to which Guidelines were met, are shown in Table 16. Twenty-eight of the 47 adapted Guidelines were rated

Table 16. Judicial Performance Program ratings, by category, using the ABA Guidelines, adapted checklist.

Category	ABA Guidelines, Adapted Checklist					
	Ratings					
	# Items	Met	Part Met	Not Met	?	Mean Rating
Goals & Purposes	4	4	0	0	0	3.0
Administration & Support	3	2	0	0	1	3.0
Performance Criteria	23	14	1	5	3	2.5
Methodology	10	7	2	0	1	2.8
Uses & Dissemination	7	1	0	0	6	—
Total	47 (100%)	28 (60%)	3 (6%)	5 (11%)	11 (23%)	2.6

as "Met" (60%). Three (6%) were rated "Partially Met" and five (11%) were rated "Not Met." As with the results of the ratings on the Standards, just under a quarter of the checklist items were not deemed appropriate for inclusion at this time (11 of 47, 23%). Here again, the primary reason for the omissions was the current deferring of performance evaluation reports. In this instance, however, a second reason for omitting additional items involved three items under the category "Performance Criteria" that appear appropriate only for the appellate level or highest administrative levels.

**Table 15. Summary of findings and recommendations, ABA Guidelines
(adapted in checklist form)**

Checklist, ABA Guidelines	Addressed	Met = 3 Part Met = 2 Not Met = 1	Findings/Comments
Goals and Purposes (GP)			
GP-1. Are goals and purposes clear and supportive of performance improvement as the primary use of evaluation?	Yes	3	
GP-2. Are goals and purposes consistent with sound judicial principles, mission, and needs of the public?	Yes	3	
GP-3. Is Program structured and implemented so it does not impair the independence of the judiciary?	Yes	3	
GP-4. Are additional (secondary) purposes recognized and considered appropriate by "governing" body?	Yes	3	Additional purposes are explicitly stated in Program's objectives.
Administration and Support (AS)			
AS-1. Is responsibility for program development and implementation vested in highest court?	Yes	3	
AS-2. Is day-to-day implementation monitored by broad-based group of individuals representing judges, lawyers, and non-lawyers familiar with the judicial system?	Yes	3	Special Committee on Judicial Performance
AS-3. Is program adequately funded and staffed?	Yes	?	Dependence on court staff for case selection; additional resources may be required; needs follow-up study.
Performance Criteria (PC)			
PC-1. Are judges evaluated on integrity, including:			
PC-1a. Avoidance of impropriety?	No	1	No items.

Checklist, ABA Guidelines	Addressed	Met = 3 Part Met = 2 Not Met = 1	Findings/Comments
PC-1b. Freedom from personal bias?	Yes	3	Scale 3:#6
PC-1c. Decisions on issues not influenced by external pressures?	Yes	3	Functionally identical to "impartiality." (See next item)
PC-1d. Impartiality?	Yes	3	Scale 3:#6-8
PC-2. Are judges evaluated on knowledge, understanding, and execution of the law, including:			
PC-2a. Issuance of legally sound decisions?	Yes	2	Scale 1:#6. (Item serves as an indirect measure only.)
PC-2b. Substantive, procedural, and evidentiary law?	Yes	3	Scale 1:#1-3
PC-2c. Factual and legal issues before the court?	Yes	3	Scale 1:#4,5,8
PC-2d. Application of judicial precedents, and other sources of authority?	Yes	3	Scale 1,#5
PC-3. Are judges evaluated on communication skills, including:			
PC-3a. Clarity of bench rulings and other oral communications?	Yes	3	Scale 1:#7

Checklist, ABA Guidelines	Addressed	Met = 3 Part Met = 2 Not Met = 1	Findings/Comments
PC-3b. Quality of written opinions; clarity and logic?	Yes	3	Scale 1: #9,10
PC-3c. Sensitivity to one's impact relating to demeanor, non-verbal communication?	Yes	3	Scale 3
PC-4. Are judges evaluated on judicial management skills, including:			
PC-4a. Preparation, attentiveness, and control over proceedings?	Yes	3	Scale 2: #1-3
PC-4b. Devoting appropriate time to all pending matters?	Yes	3	Scale 2: #1,4,5
PC-4c. Discharging administrative responsibilities diligently?	No	1	No items.
PC-4d. Management of calendar (e.g., number, age, and status of pending cases)?	No	1	No items.
PC-4e. Punctuality (i.e., prompt disposition of pending matters while maintaining rules of court)?	Yes	3	Scale 2: #1,4
PC-5. Are judges evaluated on courtroom demeanor, including:			
PC-5a. Courtesy?	Yes	3	Scale 3: #2

Checklist, ABA Guidelines	Addressed	Met = 3 Part Met = 2 Not Met = 1	Findings/Comments
PC-5b. Willingness to permit parties to be heard?	Yes	3	Scale 2:#5 Scale 3:#4,7,8
PC-6. Are judges evaluated on service to profession and to public, including:			
PC-6a. Participation in judicial education programs?	No	1	No items.
PC-6b. Assurance to the public that members of the judiciary serve to the best of their ability?	No	1	No items.
PC-7. Are judges evaluated on effective working relationships with other judges, including:			
PC-7a. Exchange of ideas, opinions with other judges (when party of a multi-judge panel)?	No	--	Appears appropriate for appellate level.
PC-7b. Sound critiquing of colleagues' work?	No	--	Appears appropriate for appellate level.
PC-7c. Facilitating performance of other judges' administrative responsibilities?	No	--	Appears appropriate for appellate and highest administrative levels.
Methodology			
M-1. Was Program developed systematically?	Yes	3	

Checklist, ABA Guidelines	Addressed	Met = 3 Part Met = 2 Not Met = 1	Findings/Comments
M-2. Is program execution sufficiently flexible, allowing improvements to be made?	Yes	3	
M-3. Is Program periodically assessed?	Yes	3	
M-4. Are appropriate, additional criteria for performance evaluation developed for use in jurisdictions (courts) that have unique characteristics and specific needs?	No	2	Not totally fitting for district court or family court, but useful.
M-5. Does the evaluation process include data collection, synthesis/analysis, and usage?	Yes	3	But "usage" has yet to really occur.
M-6. Are the analyses, evaluation timetable, and use of evaluation results appropriate for type of jurisdiction and extent of judges' experience?	Yes	?	Use has yet to occur; schedule looks reasonable.
M-7. Are methods for data collection and analysis developed with assistance with (measurement) experts to ensure quality?	Yes	3	
M-8. Are reliable, valid and multiple sources of evaluation information employed?	Yes	2	Lacks multiple sources; validity likely ok for low-stakes uses only.
M-9. Is information on performance based on "personal and current knowledge?"	Yes	3	
M-10. Are provisions for confidentiality (of judges' ratings and respondents' identities) established?	Yes	3	Security measures strong.

Checklist, ABA Guidelines	Addressed	Met = 3 Part Met = 2 Not Met = 1	Findings/Comments
Uses and dissemination			
UD-1. Is the dissemination of results consistent with Program's purpose?	Yes	?	Tentative plans seem okay.
UD-2. Is confidentiality of data and results maintained?	Yes	3	All indications are that this is given very careful attention.
UD-3. Are results shared with individual participating judges as well as senior or administrative judges?	No	?	Not implemented as yet.
UD-4. Are unwarranted and potentially misleading analyses of individual judges avoided?	No	?	Not implemented as yet.
UD-5. If additional uses of evaluation results are required, are results provided to responsible parties without the promotion of a particular philosophy?	No	?	Not implemented as yet.
UD-6. If additional uses of evaluation results are required, are results provided to responsible parties after judges are afforded an opportunity to review and comment on the results?	No	?	Not implemented as yet.
UD-7. Does the use of performance evaluation exclude the use of results for purposes of discipline?	No	?	Not implemented as yet.

All five of the "Not Met" ratings fell in the Performance Criteria category. Items addressing these five specific areas of performance evaluation were not found in the *Lawyer's Questionnaire*. Perhaps most important of these is the absence of item(s) related to the "avoidance of impropriety." Examples of other performance criteria not addressed nor considered even partially met include, "discharging administrative responsibilities diligently," and "management of calendar."

Mean (average) ratings were computed for each of the five broad categories and for the overall total. Based on a 3-point scale, items that were "Met" were assigned a value of "3." Items "Partially Met" were assigned a "2," and items "Not Met" were scored "1." Thus the highest possible mean rating was a 3.0 and the lowest possible was 1.0.

The mean ratings ranged from 2.5 for Performance Criteria, to a high of 3.0 for both Goals and Purposes, and Administration and Support. The mean rating overall was 2.6. Computing a mean rating for Uses and Dissemination was not possible nor meaningful since six of the seven items were omitted.

How does the Program stack up against the ABA Guidelines? Quite well, considering that like the application of the Standards, a number of items are excluded because those functions (e.g., "... results shared with individual participating judges ...") have not yet been implemented. Sixty per cent of all 47 items were rated as "Met." The five items rated "Not Met," as mentioned above, were all specific performance criteria recommended by the ABA for inclusion in judicial performance evaluations. None are considered "fatal flaws."

As with the 21 Standards, the ratings on the 47 Guidelines checklist items suggest that the Program's evaluation process overall is basically sound.

(3) *Correspondence between Standards and Guidelines*

The combined use and independent application of the Personnel Evaluation Standards and the ABA Guidelines has made the evaluation of the Program more comprehensive than first proposed. Originally, the intent of the present evaluation was to apply the Personnel Standards in singular fashion, to assess, generally, the soundness of the Program's procedures.

The Standards and the Guidelines are compatible conceptually. The Personnel Standards tend to be general in nature, with content relevant to virtually any personnel evaluation effort. The ABA Guidelines, in contrast, are specifically tied to personnel performance in the judiciary. Thus it was of more than passing interest to learn more about the correspondence between the two.

Appendix H is an item-by-item comparison of the checklist adapted from the Guidelines with the Personnel Evaluation Standards. The format of this comparison lists first, each of the 47 items in the checklist on the left. Then, for each item in the checklist, a corresponding item among the 21 Standards was listed to the right, if an appropriate, albeit more general counterpart, was found.

Using this method, the results reflect a substantial overlap between the two. Roughly 80% of the items on the ABA Guidelines checklist were at least partially addressed by similar domains or criteria developed for the Personnel Standards (38 of 47 items). Within the 80% or 38 items that do have Standards counterparts, all 21 Standards are accounted for, at least to some degree. Also, the Standards can and do pertain to more than one Guideline item. The opposite is true as well. Several Guideline items corresponded to more than a single Standard.

So the Standards and the Guidelines, in turn, are at times more expansive, and at other times more inclusive than the other. And, despite the Standards being broader in concept and in scope than the Guidelines, there was a surprisingly strong correspondence found between the two. In the final analysis, assessing the validity of the Judicial Performance Program's evaluation process both from the perspective of personnel evaluations generally, and from judicial performance evaluation specifically, provided a useful, convergent approach to the validation of the Program's process.

C. Data Interpretation and Presentation

There are two types of distinctly different levels of information needed from the Judicial Performance Program. First, for purposes of judicial self-improvement, individual judges need to have a profile of their evaluation results. Such a profile should contain summary information for at least 20 or more respondents that provides descriptive statistics for items and for each of the three scales (Legal Ability, Judicial Management Skills, Comportment). Frequency tabulations showing the percent of responses made for each rating value per item would provide considerable detail that would serve to highlight relative strengths and weaknesses at the item level. In addition, mean values, for both items and scales, would provide overall indications of perceived performance levels. In future consultation with program staff, we will discuss possibilities for enhancing reports, such as the use of graphic displays, and automating production through the use of data linkages and scripting software.

Second, an aggregate summary that could be used to provide a snapshot of judges' performance collectively is recommended. The summary could be used for designing judicial education activities, and, possibly, might be used for public release. The summary we envision would provide a statewide breakdown of the number of judges for Circuit, District, and Family Courts, separately, whose average ratings for Legal Ability, Judicial Management Skills, and Comportment fall within the following intervals: 1.0 - 1.4, 1.5 - 2.4, 2.5 - 3.4, 3.5 - 4.4, and 4.5 - 5.0. Such a summary would not identify the individual judges involved. It would have the distinct advantage of not encouraging unfair comparisons. Also, it would not result in attributions of excess precision to the mean ratings. And, it would summarize results in a way that relates directly back to the basic rating scale: "Poor," "Less than Adequate," "Adequate," "Good," and "Excellent."

Narrative comments that respondents may supply should be "sanitized" to remove any personally identifying information, and provided *only* in profiles that are distributed to the judge involved and the Chief Justice. Such comments should not be included in the aggregate statewide summary.

IV. Conclusions

Data analysis conducted with Judicial Performance Program questionnaire records for 12 judges from the Circuit Court of the First Judicial Circuit found that, at the level of the individual lawyer-respondent, the *Lawyer's Questionnaire* produces highly reliable information. Reliability estimates (Cronbach's alpha coefficient) were .97, .94, and .96 for the Legal Ability, Judicial Management Skills, and Comportment scales, respectively. Limitations on the precision of mean scores per judge are due largely to sampling error, that is, the sampling variation introduced via the particular set of attorneys who happen to be surveyed. As a rough rule-of-thumb, means for each of the three scales would have to be different by at least 0.5 of a scale point (e.g., half the distance between "Good" and "Excellent" on the rating scale), for samples of 20 respondents, before one might conclude that a statistically significant mean difference had been found.

Several questionnaire items were found that had relatively large rates of 'Not Applicable' (NA) response. Follow up analyses identified seven items (Legal Ability, #3, 6, 11; Judicial Management Skills, #6, 7, 8; Comportment, #3) for which the frequency of NA response was related to the type of proceeding. Motions, as contrasted with trials, incurred higher rates of NA response, suggesting that the applicability of the questionnaire's content may be limited by the type of proceeding. (Partial confirmation of these findings was obtained from analysis of a District Court data file.)

Other data analyses explored the possible concern that the occurrence of NA responses and omitted items (blanks), which are treated as missing data values when computing means for each of the three scales, might bias the judge's summary results. Generally, it did not appear that the occurrence of NA responses and omitted items (blanks) was associated with any systematic increase or decrease in mean ratings for individual judges.

A final series of exploratory analyses examined the variation in mean scores in relation to attorney-respondent type (prosecutor/plaintiff v. defense), type of proceeding (civil v. criminal), and outcome of the proceeding (trial won v. lost, motion granted v. denied). Some statistically significant differences were found. It was concluded that, if Planning and Evaluation Division staff select or reject survey records for inclusion in profiles (e.g., in order to eliminate "duplicates"), that ought to be done with the knowledge that the type of proceeding (civil or criminal), and outcome of the proceeding matter in terms of mean scores. If selection is necessary and if choices are available, selection procedures should attempt to maintain the representativeness of type and outcomes of the proceedings typical of *each* judge.

Interviews conducted with a total of 16 administrative and senior judges as well as members of the Special Committee on Judicial Performance clearly indicated that the Judicial Performance Program does *not* negatively interfere with normal courtroom practices, nor is it perceived currently as infringing upon the independence or integrity of the judiciary. In terms of purposes of the Judicial Performance program, the 16 interviewees endorsed as

appropriate "low stakes" uses, self-improvement and judicial education, without any qualifying comments. While all purposes, with the exception of promotion, were largely endorsed as "appropriate," the high stakes purposes (assignment, promotion, and retention) were sometimes given qualified endorsements.

Of the 13 interviewees who were able to supply a quantitative response to the question about the extent to which the questionnaire covered the responsibilities of a judge, all responses were in the range of 75% to 100%, and ten (10) of these 13 interviewee's responses corresponded to a coverage of at least 90%. While it appears that most interviewees viewed the *Lawyer's Questionnaire* as covering almost completely the duties and responsibilities of a judge, there is some question about which kinds of proceedings, and judges, the instrument is appropriate. Specifically, the fit of the *Lawyer's Questionnaire* and program procedures with district court proceedings, especially those in the small Neighbor Island courts, may be poor.

The Judicial Performance Program's evaluation process was found to be basically sound when the assumed purpose was that of improving individual judges' performances through evaluative feedback to the Chief Justice and to participating judges. The present program evaluation applying both the Joint Committee's Personnel Evaluation Standards and the American Bar Association's Guidelines for the Evaluation of Judicial Performance found the Program to be in good standing overall.

The Program held up quite well against the general, exemplary requirements of the Standards and the ABA Guidelines which are more specific to the judiciary. Importantly, the ratings on the Standards were strong in areas such as measurement reliability, validity and evaluator credibility.

One area of some uncertainty is that of case selection. It is not known at the present time whether potential conflict of interest is problematic in the process of court staff determining which cases constituted "meaningful opportunities."

The ultimate appraisal of the validity of the Program's evaluation process rests heavily on the eventual use and dissemination of performance evaluation results. It appears to have been a good move to have piloted and phased-in the early implementation of the Program. The decision to defer reporting of performance evaluation results for individual judges until the soundness of the Program overall could be assessed also seems to be a wise decision.

The Program's administration and support appears to be quite adequate. There is good overall monitoring of data collection and procedures. Access to files are adequately controlled and confidentiality is not compromised. Documentation is adequate especially with respect to instruction sheets for mail-outs, but the Policies and Procedures Manual should be updated and its organization improved.

Procedures outlined seem to ensure that what is planned is what actually occurs. The revised Optical Mark Reader scannable *Lawyer's Questionnaire* is a noticeable improvement in both form and function. What the Program does seem to lack is an overall long-range plan.

V. Recommendations

Instrument Validity

- 1) Consider adding an item to the *Lawyer's Questionnaire* that taps "avoidance of impropriety" in the Comportment section.
- 2) Revise item #2 in the "Background Characteristics" section of the *Lawyer's Questionnaire*. The item, which asks for number of years that the respondent practiced law, has overlapping intervals.

Validity of Evaluation Process

- 1) Examine more carefully the present procedure used for case selection; specifically, the operational definition of "meaningful opportunity," possibly by considering a dual requirement of substantive content and a minimum time duration.
- 2) Study carefully the release of results for secondary purposes such as retention. There is potential here for conflict with the purposes of self-improvement.
- 3) Consider issuing a policy statement regarding the sanctioned uses and dissemination of Program results.
- 4) Continue to keep case and court identifying information (Judicial Survey Information Form) separate from the *Lawyer's Questionnaire*. Current practices serve to protect the confidentiality of respondents and should be continued.
- 5) When the procedure allowing unmatched questionnaires is implemented, consider conducting a study to determine what effect, if any, it has on a judge's mean ratings.
- 6) Consider developing and testing a "high volume and special proceedings" lawyer's questionnaire for District Court proceedings.
- 7) Continue to use the requirement of a minimum of 20 questionnaires for producing a judicial performance profile, but remove the requirement of 20 non-blank responses per item.
- 8) Explore the possible merit of issuing a periodic report to court staff that summarizes the number of questionnaires received for participating judges that will aid them in monitoring progress on data collection efforts. (An accompanying cover memo could reaffirm the availability of technical assistance that might be needed.)

Data Interpretation & Presentation

- 1) Narrative comments that respondents may supply should be "sanitized" to remove any personally identifying information, and provided *only* in profiles that are distributed to the judge involved and the Chief Justice. Such comments should not be included in other reports.
- 2) An aggregate summary that could be used to provide a snapshot of judges' performance collectively is recommended. The summary could be used for designing judicial education activities, and, possibly, might be used for public release.
- 3) Profiles for self-improvement use by individual judges should provide descriptive statistics for items and for each of the three scales (Legal Ability, Judicial Management Skills, and Comportment).

REFERENCES

- American Bar Association, Special Committee on Evaluation of Judicial Performance. (August 1985). *American Bar Association Guidelines for the Evaluation of Judicial Performance*. Washington, DC: American Bar Association.
- American Bar Association, National Project on Judicial Performance Evaluation. (August 1989). *Judicial Performance Evaluation Planning*. Washington, DC: American Bar Association.
- American Bar Association, National Project on Judicial Performance Evaluation. (June 1990). *Supplement to the August, 1989 Judicial Performance Planning Report: Trends in Judicial Performance Assessment*. Washington, DC: American Bar Association.
- Committee on Judicial Evaluation. (November 1986). *Recommendations of the Study Committee on Judicial Evaluation*. Report to the Hawaii State Judicial Conference, November 1986. Honolulu, HI: State of Hawaii, The Judiciary.
- Feldt, L. S. and Brennan, R. L. Reliability. [Chapter 3.] (1989). In Robert L. Linn (ed.), *Educational Measurement* (3rd ed.). Washington, D.C.: National Council for Measurement in Education and American Council on Education.
- Keilitz, S. and McBride, J. W. (Winter 1992). Judicial performance evaluation comes of age. *State Court Journal*, 16(1), 4-13.
- Nunnally, J. C. (1978). *Psychometric Theory*. (2nd ed.). New York, N.Y.: McGraw-Hill.
- Revised chart for "Judicial performance comes of age." (Vol. 16, No. 1, Winter 1992). (Summer 1992). *State Court Journal*, 16(3), 30-33.
- State of Connecticut, Judicial Branch. (1991). *Judicial Performance Program. Annual Report to the Chief Court Administrator*. Hartford, CT: Judicial Branch, Judicial Evaluation Administrator.
- State of Hawaii, The Judiciary. (May 1994). *Report on the Judicial Performance Program*. Honolulu, HI: State of Hawaii, The Judiciary.
- State of Hawaii, The Judiciary. (1994). "Judicial Performance Program Policies and Procedures." Honolulu, HI: State of Hawaii, The Judiciary. [Policies & procedures manual]
- State of Hawaii, Supreme Court. (1991). *Rule 19. Judicial Performance Program*. Honolulu, HI: State of Hawaii, The Judiciary.

Supreme Court of New Jersey. (June 2, 1988). *Rule 1:35A. Judicial Performance Program*. Trenton, NJ: Committee on Judicial Performance.

Supreme Court of New Jersey, Committee on Judicial Performance. (no date). "Appellate Judges' Questionnaire;" "Lawyers' Questionnaire;" "Lawyers' Questionnaire Relating to High Volume Courts and Special Proceedings;" "Juror's Questionnaire" [includes cover letters] Trenton, NJ: Committee on Judicial Performance.

Supreme Court of New Jersey, Committee on Judicial Performance. (December 1989). *First Report on the Judicial Performance Program*. Trenton, NJ: The New Jersey Supreme Court, Committee on Judicial Performance.

Appendix A: *Lawyer's Questionnaire* (Sample Copy)



































































Record No.

 USE NO. 2 PENCIL ONLY

0	0	0	0	0	0	0
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9

[For Office Use Only]

THIS SECTION DEALS WITH LEGAL COMPETENCE, LEARNING, AND UNDERSTANDING. IT ALSO DEALS WITH THE JUDICIAL APPLICATION OF KNOWLEDGE IN THE CONDUCT OF COURT PROCEEDINGS.

1. KNOWLEDGE OF RELEVANT SUBSTANTIVE LAW	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
2. KNOWLEDGE OF RULES OF PROCEDURE	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
3. KNOWLEDGE OF RULES OF EVIDENCE	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
4. ABILITY TO IDENTIFY AND ANALYZE RELEVANT ISSUES	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
5. JUDGEMENT IN APPLICATION OF RELEVANT LAWS AND RULES	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
6. GIVING REASONS FOR RULINGS WHEN NEEDED	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
7. CLARITY OF EXPLANATION OF RULINGS	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
8. ADEQUACY OF FINDINGS OF FACT	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
9. CLARITY OF JUDGE'S DECISION (ORAL/WRITTEN)	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
10. COMPLETENESS OF JUDGE'S DECISION (ORAL/WRITTEN)	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 
11. JUDGE'S CHARGE TO THE JURY	EXCELLENT 	GOOD 	ADEQUATE 	LESS THAN ADEQUATE 	POOR 	NOT APPLICABLE 

SAMPLE

LAWYER'S QUESTIONNAIRE (Circuit Court)

THIS IS PAGE 2 OF A 4-PAGE QUESTIONNAIRE.

SECTION 2: JUDICIAL MANAGEMENT SKILLS

THIS SECTION DEALS WITH JUDICIAL ABILITY AND SKILL
IN THE ORGANIZATION, MANAGEMENT, AND HANDLING OF
COURT PROCEEDINGS.

USE NO. 2 PENCIL ONLY

MARK ONLY ONE RESPONSE FOR EACH QUESTION, AND
PLEASE BE SURE TO ANSWER EVERY QUESTION.

1. MOVING THE PROCEEDING IN AN APPROPRIATELY EXPEDITIOUS MANNER	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
2. MAINTAINING PROPER CONTROL OVER THE PROCEEDING	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
3. DOING THE NECESSARY HOMEWORK ON THE CASE	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
4. RENDERING RULINGS AND DECISIONS WITHOUT UNNECESSARY DELAY	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
5. ALLOWING ADEQUATE TIME FOR PRESENTATION OF THE CASE IN LIGHT OF EXISTING TIME CONSTRAINTS	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
6. RESOURCEFULNESS AND COMMON SENSE IN RESOLVING PROBLEMS ARISING FROM THE PROCEEDING	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
7. SKILLS IN EFFECTING COMPROMISE	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
8. INDUSTRIOUSNESS	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>

SAMPLE

LAWYER'S QUESTIONNAIRE (Circuit Court)

THIS IS PAGE 3 OF A 4-PAGE QUESTIONNAIRE.

SECTION 3: COMPORTMENT

THIS SECTION DEALS WITH VARIOUS ASPECTS OF JUDICIAL PERSONALITY AND BEHAVIOR IN THE COURT PROCEEDINGS, SUCH AS TEMPERAMENT, ATTITUDE, AND MANNER.
MARK ONLY ONE RESPONSE FOR EACH QUESTION, AND PLEASE BE SURE TO ANSWER EVERY QUESTION.



1. ATTENTIVENESS	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
2. COURTESY TO PARTICIPANTS	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
3. COMPASSION	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
4. PATIENCE	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
5. ABSENCE OF ARROGANCE	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
6. ABSENCE OF BIAS AND PREJUDICE BASED ON RACE, SEX, ETHNICITY, RELIGION, SOCIAL CLASS, OR OTHER FACTOR	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
7. EVENHANDED TREATMENT OF LITIGANTS	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>
8. EVENHANDED TREATMENT OF ATTORNEYS	EXCELLENT <input type="radio"/>	GOOD <input type="radio"/>	ADEQUATE <input type="radio"/>	LESS THAN ADEQUATE <input type="radio"/>	POOR <input type="radio"/>	NOT APPLICABLE <input type="radio"/>

SAMPLE LAWYER'S QUESTIONNAIRE (Circuit Court)

THIS IS PAGE 4 OF A 4-PAGE QUESTIONNAIRE.

SECTION 4: BACKGROUND CHARACTERISTICS

THIS INFORMATION WILL BE USED FOR STATISTICAL
PURPOSES ONLY.



1. HOW MANY TIMES HAVE YOU APPEARED BEFORE THIS JUDGE IN OTHER
PROCEEDINGS IN THE PAST THREE YEARS?

0 TO 5 6 TO 10 11 TO 15 16 TO 20 21 TO 25 MORE THAN 25 TIMES

2. HOW LONG HAVE YOU PRACTICED LAW? (YEARS)

0 TO 3 3 TO 6 6 TO 9 9 TO 12 12 TO 15 15 TO 18 18 TO 21 21 TO 24 MORE THAN 24

3. WHICH OF THE FOLLOWING DESCRIBES YOUR PRACTICE OF LAW?

☐ SOLO (INCLUDING OFFICE SHARING)

☐ LAW FIRM WITH 2-15 ATTORNEYS

☐ STATE PUBLIC DEFENDER

☐ LAW FIRM WITH MORE THAN 15 ATTORNEYS

☐ CITY PROSECUTOR

☐ CORPORATE OR HOUSE COUNSEL

☐ STATE ATTORNEY GENERAL

☐ CITY CORPORATION COUNSEL

☐ PRO SE (REPRESENTING SELF)

☐ OTHER (PLEASE SPECIFY)

4. HOW WOULD YOU DESCRIBE THE OUTCOME OF THE SELECTED PROCEEDING?
(CHOOSE ONE TYPE OF PROCEEDING, AND PLEASE FILL IN ONE OVAL.)

TRIAL: ☐ WON ☐ WON IN PART ☐ LOST ☐ DISMISSED ☐ OTHER

MOTION: ☐ GRANTED ☐ DENIED ☐ WITHDRAWN ☐ OTHER

SAMPLE

COMMENT PAGE (Circuit Court)

IF YOU WISH TO ELABORATE ON ANY OF YOUR RESPONSES OR COMMENT FURTHER ON ASPECTS OF THE JUDGE'S PERFORMANCE OR IF YOU HAVE ANY COMMENTS ON THE CONTENT OR FORM OF THIS QUESTIONNAIRE AND WAYS IN WHICH IT COULD BE IMPROVED, PLEASE USE THE SPACE BELOW. PLEASE TYPE YOUR COMMENTS, AND REMEMBER NOT TO IDENTIFY YOURSELF.

Appendix B: Additional Data Analysis Summary

All findings reported below are from analysis of Circuit Court of the First Circuit data unless noted otherwise. The two data files analyzed, the "Duplicated Data File" and the "Unduplicated Data File" have been previously described in the main body of the report.

Results of the analyses are organized below into the following sections: (a) Characteristics of the Proceedings, (b) Characteristics of the Respondents, and (c) Items Relating to the Survey.

(a) Characteristics of the Proceedings

Characteristics of the proceeding provide some context information to describe the setting for the judicial evaluation via the *Lawyers' Questionnaire*. The type of proceeding, categorized as civil or criminal, had been extracted from the "Judicial Survey Information Form" and included in the survey data files. The following table shows the breakdown of survey records for this variable. The ratio of civil-to-criminal proceedings was 3-to-2.

Type of Proceeding	Duplicated Data File	Unduplicated Data File
Civil	58%	60%
Criminal	42%	40%
N (# respondents)	311	287

The last structured item of the questionnaire (question "4" or Q4 of "Section 4: Background Characteristics") asked the respondent to describe the outcome of the proceeding. As shown in the following table, less than one-third of the proceedings were trials. The bulk of the survey information (70%) originate from an evaluation setting of proceedings that are motions, at least for these data files for the Circuit Court of the First Judicial Circuit.

Outcome of Proceeding	Duplicated Data File	Unduplicated Data File
Trial Won	10%	10%
Trial Won in Part	7%	7%
Trial Lost	7%	7%
Trial Dismissed	1%	1%
Trial Other	4%	4%
Subtotal TRIAL	29%	30%
Motion Granted	38%	38%
Motion Denied	24%	22%
Motion Other	9%	10%
Subtotal MOTION	71%	70%
N (# respondents)	305	284
# missing (blank)	6	3

Note: Column totals may vary slightly from the subtotals indicated due to rounding.

(b) Characteristics of the Respondents

Data obtained from the "Judicial Survey Information Form" and included in the data files categorized the attorney-respondent as prosecutor/plaintiff or defense. As shown in the following table, the respondents comprise nearly equally-sized groups on this characteristic.

Attorney-Respondent Type	Duplicated Data File	Unduplicated Data File
Prosecutor/Plaintiff	47%	46%
Defense	53%	54%
N (# respondents)	311	287

The first three items of "Section 4: Background Characteristics" ask the respondent: (Q1) how many times they "have appeared before this judge in other proceedings in the past three years;" (Q2) how long they have practiced law (in years); and, (Q3) from a given set of categories, to identify their type of law practice. The following tables show the characteristics of the respondents in terms of these three variables.

# Times Before This Judge?	Duplicated Data File	Unduplicated Data File
0 to 5	48%	49%
6 to 10	15%	16%
11 to 15	7%	7%
16 to 20	9%	8%
21 to 25	2%	2%
More than 25	20%	17%
N (# respondents)	293	271
# missing (blank)	18	16

Note: Column totals may vary slightly from 100% due to rounding.

# Years Practiced Law?	Duplicated Data File	Unduplicated Data File
0 < 3	5%	6%
3 < 6	14%	14%
6 < 9	20%	19%
9 < 12	17%	17%
12 < 15	11%	11%
15 < 18	8%	8%
18 < 21	13%	12%
21 < 24	5%	4%
25 or more	7%	8%
N (# respondents)	309	286
# missing (blank)	2	1

Note: Column totals may vary slightly from 100% due to rounding.

Type of Law Practice?	Duplicated Data File	Unduplicated Data File
Solo	21%	21%
State Public Defender	13%	13%
City Prosecutor	19%	18%
Law Firm (w. 2 to 15)	29%	30%
Law Firm (w. 15 or more)	15%	16%
Others	3%	3%
N (# respondents)	309	286
# missing (blank)	2	1

Note 1: "Others" is a composite category formed by combining "State Attorney General," "Corporate or House Counsel," and "City Corporation Counsel."

Note 2: Column totals may vary slightly from 100% due to rounding.

(c) Items Relating to the Survey

The data files contained two data elements pertaining to the survey itself: the date it was returned to the Planning and Evaluation Division, and whether or not the respondent had provided written comments on the final "Comment Page" of the survey form. As shown in the following two tables, the return of the surveys spanned nearly 13.5 months, and about one-third of the respondents provided written comments.

Date Survey Returned	Duplicated Data File	Unduplicated Data File
Beginning Date	10/1/93	10/1/93
Ending Date	12/12/94	12/12/94
N (# respondents)	311	287

Written Comments Made?	Duplicated Data File	Unduplicated Data File
No	70%	68%
Yes	30%	32%
N (# respondents)	311	287

Appendix C: Interview Protocols

Interview Protocol -- Administrative Judges and Senior Judges

Judicial Performance Program External Evaluation

Judge: _____

Court: _____

Phone: _____

Interviewed by: _____

Contact Attempts -- Date: _____

Time: _____

[Introduction: Self-identification; purpose of interview; confidentiality]

1. Prior to the implementation of Hawaii's Judicial Performance Program, what process, if any, was used to evaluate judges?
2. At the present time, in addition to the Judicial Performance Program, is any other formal evaluation process used to regularly evaluate judges? (If "Yes," ask what it is, i.e., what it's called, who does it, get a brief description or explanation of the process.)
No ____ Yes ____ (follow-up):
3. Does the Judicial Performance Program's implementation interfere with normal courtroom practices? (If "Yes," ask in what way(s) it interferes.)
No ____ Yes ____ (follow-up):
4. Does the Judicial Performance Program infringe upon the independence or integrity of the judiciary? (If "Yes," ask how or in what ways it infringes.)
No ____ Yes ____ (follow-up):

5. The items in the *Lawyers' Questionnaire* used in the Judicial Performance Program cover three areas: Legal Ability, Judicial Management Skills, and Comportment. Considering all that you are responsible for as a judge, on a scale of 0% to 100%, what percent of your responsibilities would you say are covered by the questionnaire? (For responses less than 100%, ask [a] what areas of responsibility are missing and [b] how each missing area compares in importance -- more, equal, less -- to those on the questionnaire.)

	_____ %	
Area missing:	_____	More Equal Less
Area missing:	_____	More Equal Less
Area missing:	_____	More Equal Less

6. Following are five purposes for which judicial evaluation programs similar to Hawaii's have been used. I will read each purpose, one by one. Please identify whether you think the purpose is appropriate or inappropriate. If you deem the purpose inappropriate, please tell me why.

	OK	Not OK (follow-up):
(a) self-improvement	_____	_____
(b) designing judicial education programs	_____	_____
(c) assignment of judges	_____	_____
(d) promotion decisions	_____	_____
(e) retention decisions	_____	_____

7. Are there any notable strengths or weaknesses of the Judicial Performance Program you would like me to note?

[Closing & "Thank You!"]

(intview.1a; 1/18/95)

Interview Protocol -- Special Committee on Judicial Performance

Judicial Performance Program External Evaluation

Committee Member: _____

Title: _____

Phone: _____

Interviewed by: _____

Contact Attempts -- Date: _____

Time: _____

[Introduction: Self-identification; purpose of interview; confidentiality]

1. At the present time, in addition to the Judicial Performance Program, is any other formal evaluation process used to regularly evaluate judges? (If "Yes," ask what it is, i.e., what it's called, who does it, get a brief description or explanation of the process.)
No ____ Yes ____ (follow-up):

[Questions #2 and #3 should be asked only of judges or attorneys.]

- *2. Does the Judicial Performance Program's implementation interfere with normal courtroom practices? (If "Yes," ask in what way(s) it interferes.)
No ____ Yes ____ (follow-up):

- *3. Does the Judicial Performance Program infringe upon the independence or integrity of the judiciary? (If "Yes," ask how or in what ways it infringes.)
No ____ Yes ____ (follow-up):

4. The *Lawyers' Questionnaire* used in the Judicial Performance Program covers three areas of judicial responsibility: Legal Ability, Judicial Management Skills, and Comportment. Considering all of a judge's responsibilities and duties as totaling 100%, on a scale of 0% to 100%, what percent of judges' responsibilities would you say are covered by the questionnaire? (For responses less than 100%, ask [a] what areas of responsibility are missing and [b] how each missing area compares in importance -- more, equal, less -- to those on the questionnaire.)

	_____ %	
Area missing:	_____	More Equal Less
Area missing:	_____	More Equal Less
Area missing:	_____	More Equal Less

5. Following are five purposes for which judicial evaluation programs similar to Hawaii's have been used. I will read each purpose, one by one. Please identify whether you think the purpose is appropriate or inappropriate. If you deem the purpose inappropriate, please tell me why.

	OK	Not OK (follow-up):
(a) self-improvement	_____	_____
(b) designing judicial education programs	_____	_____
(c) assignment of judges	_____	_____
(d) promotion decisions	_____	_____
(e) retention decisions	_____	_____

6. Are there any notable strengths or weaknesses of the Judicial Performance Program you would like me to note?

[Closing & "Thank You!"]

(intview.2; 1/18/95)

Interviewee List

Administrative Judges and Senior Judges

Honorable Marie Milks, Administrative Judge Criminal Division, Circuit Court, First Circuit

Honorable Herbert K. Shimabukuro, Administrative Judge Civil Division, Circuit Court,
First Circuit

Honorable Michael A. Town, Senior Judge, Family Court, First Circuit

Honorable James H. Dannenberg, Administrative Judge, District Court, First Circuit

Honorable E. John McConnell, Administrative Judge & Senior Judge, Family Court, Second
Circuit

Honorable John T. Vail, Administrative Judge, District Court, Second Circuit

Honorable Ronald Ibara, Administrative Judge & Senior Judge, Family Court, Third Circuit

*Honorable Jeffrey Choi, Administrative Judge, District Court, Third Circuit

*Honorable George M. Masuoka, Administrative Judge, Circuit Court, Fifth Circuit

Honorable Gerald Matsunaga, Administrative Judge, District Court, Fifth Circuit

Special Committee on Judicial Performance

Judge Daniel Heely, Chairperson

For the Office of the Administrative Director of the Courts, The Judiciary

Ms. Elizabeth Kent, Vice-Chairperson

Center for Alternative Dispute Resolution, The Judiciary

Mr. Lowell Chun-Hoon, Attorney at Law
King, Nakamura & Chun-Hoon

Mr. Jeffrey S. Portnoy, Attorney at Law
Cades, Schutte, Fleming & Wright

Mr. Herbert Cornuelle
(representing the public)

Ms. Suzanne Peterson
(representing the public)

* Also members of the Special Committee on Judicial Performance

Appendix D: Interview Data Analysis

A series of 16 telephone interviews were conducted by the evaluators with ten (10) administrative judges and senior judges, all of whom were participants in the Judicial Performance Program, and eight (8) members of the Special Committee on Judicial Performance. Since two of the Special Committee members were also among the ten judges in the first group of interviewees, the total number of individual interviewees was 16.

The interviews provided an excellent opportunity to obtain information directly from the program's participants (judges) as well as the program's developers/planners (Special Committee members). The interview questions focused on issues of instrument validity and the evaluation process.

All interviews were conducted during the period January 20-31, 1995. All of the 16 targeted interviewees were contacted and interviews completed.

The following table lists the interview items used and the number of interviewees per item.

Interview Item	N
Prior to the implementation of Hawaii's Judicial Performance Program, what process, if any, was used to evaluate judges? [#1]	10
At the present time, in addition to the Judicial Performance Program, is any other formal evaluation process used to regularly evaluate judges? [#2; #1]	16
Does the Judicial Performance Program's implementation interfere with normal courtroom practices? [#3; #2*]	14
Does the Judicial Performance Program infringe upon the independence or integrity of the judiciary? [#4; #3*]	14
The items in the <i>Lawyers' Questionnaire</i> used in the Judicial Performance Program cover three areas: Legal Ability, Judicial Management Skills, and Comportment. Considering all that you are responsible for as a judge, on a scale of 0% to 100%, what percent of your responsibilities would you say are covered by the questionnaire? [Alternate wording for Special Commission member interviews: "Considering all of a judge's responsibilities and duties as totaling 100%, on a scale of 0% to 100%, what percent of judges' responsibilities would you say are covered by the questionnaire?"] [#5; #4]	16

Interview Item	N
<p>Following are five purposes for which judicial evaluation programs similar to Hawaii's have been used. I will read each purpose, one by one. Please identify whether you think the purpose is appropriate or inappropriate. If you deem the purpose inappropriate, please tell me why.</p> <p>(a) self-improvement (b) designing judicial education programs (c) assignment of judges (d) promotion decisions (e) retention decisions [#6; #5]</p>	16
<p>Are there any notable strengths or weaknesses of the Judicial Performance Program you would like me to note? [#7; #6]</p>	16

Notes.

- (1) Numbers in brackets in the column labelled "Interview Item" refer to the item number in the *Interview Protocol -- Administrative Judges and Senior Judges* (first number) and in the *Interview Protocol -- Special Committee on Judicial Performance* (second number). Copies of both instruments are given in *Appendix C*.
- (2) Questions #2 and #3 in the *Interview Protocol -- Special Committee on Judicial Performance* were asked of judges and attorneys only.

The following tables summarize the results of conducting content categorization and tabulation of the categorized responses provided to each question. For several items, and for follow-up responses to an item, an interviewee's answer may include multiple parts, corresponding to "mentions" in several categories. In such situations, the number of mentions, not the number of respondents, was tabulated.

(1a) Other prior processes used to evaluate judges? (n = 10)

# Mentions	Content Category
7	<i>None</i> . No formal judicial evaluation process was used.
4	<i>Informal feedback</i> (comments, suggestions) may be occasionally provided by peers, attorney friends, court staff, and others.
3	Although not a formal evaluation process, occasional <i>complaints</i> about judicial performance may be made to the Chief Justice, administrative judge, or administrative director.
2	Evaluative input is solicited periodically by the <i>Judicial Selection Commission</i> for making retention recommendations.
1	The <i>Court Observer Program</i> was implemented on a limited basis starting with the Circuit Court in Honolulu under Judge Wong, and was continued by Judge Town.

Note: The total number of mentions will not usually equal the number of persons interviewed.

(1b) Other current processes used to evaluate judges? (n = 16)

No	Yes	Other
11	3	2

Elaborations made on "No," "Yes," or "Other" response:

# Mentions	Content Category
4	<i>Judicial Selection Commission</i>
3	<i>Court Observer Program</i>
1	<i>informal exchange of information</i> between the Trial Judges' Association and the Hawaii State Bar Association
1	<i>peer review</i> (in early development stage)

(2) Interference with normal courtroom practices? (n = 14)

No	Yes	Other
12	1*	1**

* - "Makes us perform better!"

** - Don't know

(3) Infringement upon the independence or integrity of the judiciary? (n = 14)

No	Yes	Other
12	0	2*

* - Don't know

Narrative qualifications associated with response "No":

# Mentions	Content Category
3	Public release, media use, or HSBA use could result in infringement (and degenerate into a popularity contest)
1	Critical letters to the Judicial Selection Commission may be <i>more</i> problematic (than use of Judicial Performance Program information)
1	Assumption and understanding is that the purpose is furthering improvement

(4) Extent to which judge's responsibilities are covered by the *Lawyers' Questionnaire*? (n = 16)

# Interviewees	Response: Percent (%) Covered
3	100%
1	100%, for trial judges
1	nearly 100%
1	90-95%
2	90+ %
2	90%
1	80%
1	75+ %
1	75%
1	most
2	cannot say

Possible missing areas mentioned included: Administrative duties (2); settlement skills (1); conduct as a private citizen (1); mindfulness of the community (1); and, common sense & sense of humor (1). [Numbers in parentheses are the frequency of mention.]

Other comments: The questionnaire may not/does not have appropriate content for evaluating administrative and motions judges. (1 mention)

(5) Appropriate purposes of the Judicial Performance Program? (n = 16)

Purpose	Appropriate	Inappropriate	Other
self-improvement	16	0	0
designing judicial education programs	16	0	0
assignment of judges	13 ^A	2 ^a	1
promotion decisions	10 ^B	5 ^b	1
retention decisions	15 ^C	0	1

Note: The footnotes reference related narrative comments summarized immediately below.

Reservations or other conditional statements related to appropriate purposes (comments represent one interviewee unless otherwise indicated):

- A: Applicability of assignment is limited for some Neighbor Island courts.
- a: Depends on whether the questionnaire is valid for this purpose.
- B: Appropriate, but to a limited extent; unexpressed reservations (2 interviewees)
 Appropriate, but only if the survey is "statistically okay."
- b: The Judiciary in Hawaii has an appointment process, but no promotion process *per se*.
 (4 interviewees)
 Depends on whether the questionnaire is valid for this purpose.
- C: Appropriate, but to a limited extent; concerned about popularity contest developing.
 The Judicial Selection Commission must preserve the confidentiality of any
 information provided.
 Some unexpressed reservation
 Appropriate, but only if the survey is "statistically okay."

(6) Notable strengths or weaknesses of the Judicial Performance Program?

For this question, it seemed more appropriate to attempt to maintain the interviewee's "voice" (to the extent that is possible from written interview notes) rather than summarize their response in categories. Sometimes this helps to better understand what interviewees are really saying. All comments are paraphrased, although most are close to being direct quotes. The comments listing is loosely organized by similarity of content of the comments provided.

Strengths

- Program seems to be working well.
- Program seems to be very good.
- A good program, necessary to the judiciary.
- A good tool for judicial improvement.
- Program's purpose, to improve judges, is great!
- Most important aspect of the program is that it is a tool to help judges improve themselves.
- The program represents a positive first step by The Judiciary at evaluation and self-improvement.
- Assists judges who can work to improve for retention.
- Really want program to work. It can help judges and help the bar.
- Serves to surface complaints and criticism.
- Possible to have a broad array of response. In the past, most feedback was from those with a "bone to pick."
- Attempts to maintain a balance by using attorneys [prosecutor/plaintiff and defense] from the same case.

Great deal of development work and testing went into the program.

The questionnaire is good--it asks the right questions.

The questionnaire taps first-hand, recent experience (and does not try to rely on recall of long past events).

Quality staff--extremely conscientious!

The program staff have integrity, will protect confidentiality.

The Chief Justice and administrative judges are squarely behind the program.

Judges have been very supportive (not defensive or resistant) of the program and look forward to receiving evaluation results.

Having the program sends a positive message to the community.

Weaknesses

There is still a concern among some attorneys that the confidentiality of their responses may be breached.

Continues to be *some* concern about the confidentiality of the lawyers' responses, and uncertainty on how to best treat any narrative comments provided (e.g., "sanitize" or simply delete?).

It took so long to establish the program.

May be an unresolved due process issue, re: the right of judges to confront and respond to their evaluators.

Relationship of the judiciary and HSBA--but it has improved.

Concern about the complexity and detail necessary to maintain the program: Can the Chief Justice (and individual judges) get useful information without going through the time, expense, and complexity of the current evaluation?

No specific results/details have been released to anyone except the Chief Justice and some judges.

Not sure that a fair sampling of attorneys get the questionnaire.

Court staff handle the listing of attorneys for "meaningful cases" to be surveyed--could conceivably bias the distribution of surveys. (2 interviewees)

Need a fairly large number respondents for a good sampling; all courts here [Neighbor Island] have run out of attorneys to survey. Cannot repeatedly survey the same ones--a real problem here.

Serious concerns with "fit" of the program's design and operations for Neighbor Island district courts:

- High volume proceedings, preliminary hearings, and relatively brief trials, which comprise most of the work done, are not the types of "meaningful case" that qualify for using the program's survey. Those proceedings that do qualify are not really representative.
- Relatively small number of attorneys and the need to avoid surveying the same ones over and over is problematic.

Other Comments

I don't know [re: strengths, weaknesses]; its too early to really say.

No experience with program. (2 interviewees)

Other Comments (continued)

Program depends on responses from attorneys. Important for attorneys to complete the survey.

Input from judge's staff could be a rich source of evaluation information.

Would like to see program expand in the future to get a broader range of feedback (e.g., from appellate justices, court staff, jurors, litigants, witnesses, public observers).

Confidentiality of judicial profiles is essential. Otherwise it becomes a popularity contest which could lead to "judge shopping" and interfere with judicial independence.

The program's self-improvement emphasis can be helpful, but publication in the newspaper--as is done in some other states--is not helpful.

We could get buried in the numbers if results are publicized, so don't publish results except for an overall statewide summary. The Chief Justice can and should be involved in evaluating my performance, then following up (e.g., counseling could be helpful).

Regarding information release--should *never* be used as a scorecard in the newspaper.

Appendix E: Personnel Evaluation Standards (adapted to the judiciary)

Domain = Propriety

The Propriety Standards require that evaluations be conducted legally, ethically, and with due regard for the welfare of evaluatees and clients of the evaluations.

P-1: Service Orientation

Evaluations of judges/justices should promote sound judicial principles, fulfillment of institutional mission, and effective performance of job responsibilities, so that the legal needs of court participants, community, and society are met.

P-2 Formal Evaluation Guidelines

Guidelines for personnel evaluations should be recorded and provided to judges/justices in statements of policy, and/or personnel evaluation policy and procedure manuals, so that evaluations are consistent, equitable, and conducted in accordance with pertinent laws, rules, and ethical codes.

P-3 Conflict of Interest

Conflicts of interest should be identified and dealt with openly and honestly, so that they do not compromise the evaluation process and results.

P-4 Access to Personnel Evaluation Reports

Access to reports of personnel evaluation should be limited to individuals with a legitimate need to review and use the reports, so that appropriate use of the information is assured.

P-5 Interactions with Evaluatees

The evaluation should address evaluatees in a professional, considerate, and courteous manner so that their self-esteem, motivation, professional reputations, performance, and attitude toward personnel evaluation are enhanced or, at least, not needlessly damaged.

Domain = Utility

The Utility Standards are intended to guide evaluations so that they will be informative, timely, and influential.

U-1 Constructive Orientation

Evaluations should be constructive, so that they help the judiciary to improve the performance of judges/justices, individually and collectively, and encourage and assist those evaluated to provide excellent service.

U-2 Defined Uses

The users and the intended uses of a personnel evaluation should be identified, so that the evaluation can address appropriate questions.

U-3 Evaluator Credibility

The evaluation system should be managed and executed by persons with the necessary qualifications, skills, and competence, and evaluators should conduct themselves professionally, so that evaluation reports are respected and used.

U-4 Functional Reporting

Reports should be clear, timely, accurate, and germane, so that they are of practical value to the evaluatee and other appropriate audiences.

U-5 Follow-Up and Impact

Evaluations should be followed up, so that users and evaluatees are aided to understand the results and take appropriate actions.

Domain = Feasibility

The Feasibility Standards call for evaluation systems that are as easy to implement as possible, efficient in their use of time and resources, adequately funded, and viable from a number of other standpoints.

F-1 Practical Procedures

Personnel evaluation procedures should be planned and conducted so that they produce needed information while minimizing disruption and cost.

F-2 Political Viability

The personnel evaluation system should be developed and monitored collaboratively, so that all concerned parties are constructively involved in making the system work.

F-3 Fiscal Viability

Adequate time and resources should be provided for personnel evaluation activities, so that evaluation plans can be effectively and efficiently implemented.

Domain = Accuracy

The Accuracy Standards require that the obtained information be technically accurate and that conclusions be linked logically to the data.

A-1 Defined Role

The role, responsibilities, performance objectives, and needed qualifications of the evaluatee should be clearly defined, so that the evaluator can determine valid assessment.

A-2 Work Environment

The context in which the evaluatee works should be identified, described, and recorded, so that environmental influences and constraints on performance can be considered in the evaluation.

A-3 Documentation of Procedures

The evaluation procedures actually followed should be documented, so that the evaluatees and other users can assess the actual, in relation to intended, procedures.

A-4 Valid Measurement

The measurement procedures should be chosen or developed and implemented on the basis of the described role and the intended use, so that the inferences concerning the evaluatee are valid and accurate.

A-5 Reliable Measurement

Measurement procedures should be chosen or developed to assure reliability, so that the information obtained will provide consistent indications of the performance of the evaluatee.

A-6 Systematic Data Control

The information used in the evaluation should be kept secure, and should be carefully processed and maintained, so as to ensure that the data maintained and analyzed are the same as the data collected.

A-7 Bias Control

The evaluation process should provide safeguards against bias, so that the evaluatee's qualifications or performance are assessed fairly.

A-8 Monitoring Evaluation Systems

The personnel evaluation system should be reviewed periodically and systematically, so that appropriate revisions can be made.

Appendix F: Sample ratings worksheet, Personnel Evaluation Standards (adapted to the judiciary)

Standard	Met	Findings/Comments
P-1: Service Orientation Evaluations of judges/justices should promote sound judicial principles, fulfillment of institutional mission, and effective performance of job responsibilities, so that the legal needs of court participants, community, and society are met.		
P-2 Formal Evaluation Guidelines Guidelines for personnel evaluations should be recorded and provided to judges/justices in statements of policy, and/or personnel evaluation policy and procedure manuals, so that evaluations are consistent, equitable, and conducted in accordance with pertinent laws, rules, and ethical codes.		
P-3 Conflict of Interest Conflicts of interest should be identified and dealt with openly and honestly, so that they do not compromise the evaluation process and results.		
P-4 Access to Personnel Evaluation Reports Access to reports of personnel evaluation should be limited to individuals with a legitimate need to review and use the reports, so that appropriate use of the information is assured.		
P-5 Interactions with Evaluatees The evaluation should address evaluatees in a professional, considerate, and courteous manner so that their self-esteem, motivation, professional reputations, performance, and attitude toward personnel evaluation are enhanced or, at least, not needlessly damaged.		
U-1 Constructive Orientation Evaluations should be constructive, so that they help the judiciary to improve the performance of judges/justices, individually and collectively, and encourage and assist those evaluated to provide excellent service.		

Standard	Met	Findings/Comments
<p>U-2 Defined Uses</p> <p>The users and the intended uses of a personnel evaluation should be identified, so that the evaluation can address appropriate questions.</p>		
<p>U-3 Evaluator Credibility</p> <p>The evaluation system should be managed and executed by persons with the necessary qualifications, skills, and competence, and evaluators should conduct themselves professionally, so that evaluation reports are respected and used.</p>		
<p>U-4 Functional Reporting</p> <p>Reports should be clear, timely, accurate, and germane, so that they are of practical value to the evaluatee and other appropriate audiences.</p>		
<p>U-5 Follow-Up and Impact</p> <p>Evaluations should be followed up, so that users and evaluatees are aided to understand the results and take appropriate actions.</p>		
<p>F-1 Practical Procedures</p> <p>Personnel evaluation procedures should be planned and conducted so that they produce needed information while minimizing disruption and cost.</p>		
<p>F-2 Political Viability</p> <p>The personnel evaluation system should be developed and monitored collaboratively, so that all concerned parties are constructively involved in making the system work.</p>		
<p>F-3 Fiscal Viability</p> <p>Adequate time and resources should be provided for personnel evaluation activities, so that evaluation plans can be effectively and efficiently implemented.</p>		
<p>A-1 Defined Role</p> <p>The role, responsibilities, performance objectives, and needed qualifications of the evaluatee should be clearly defined, so that the evaluator can determine valid assessment.</p>		

Standard	Met	Findings/Comments
<p>A-2 Work Environment</p> <p>The context in which the evaluatee works should be identified, described, and recorded, so that environmental influences and constraints on performance can be considered in the evaluation.</p>		
<p>A-3 Documentation of Procedures</p> <p>The evaluation procedures actually followed should be documented, so that the evaluatees and other users can assess the actual, in relation to intended, procedures.</p>		
<p>A-4 Valid Measurement</p> <p>The measurement procedures should be chosen or developed and implemented on the basis of the described role and the intended use, so that the inferences concerning the evaluatee are valid and accurate.</p>		
<p>A-5 Reliable Measurement</p> <p>Measurement procedures should be chosen or developed to assure reliability, so that the information obtained will provide consistent indications of the performance of the evaluatee.</p>		
<p>A-6 Systematic Data Control</p> <p>The information used in the evaluation should be kept secure, and should be carefully processed and maintained, so as to ensure that the data maintained and analyzed are the same as the data collected.</p>		
<p>A-7 Bias Control</p> <p>The evaluation process should provide safeguards against bias, so that the evaluatee's qualifications or performance are assessed fairly.</p>		
<p>A-8 Monitoring Evaluation Systems</p> <p>The personnel evaluation system should be reviewed periodically and systematically, so that appropriate revisions can be made.</p>		

Appendix G: Correspondence between the ABA Guidelines (adapted in checklist form) and the Personnel Evaluation Standards

ABA Guidelines, adapted checklist	Personnel Standards
Goals and Purposes (GP)	
GP-1. Are goals and purposes clear and supportive of performance improvement as the primary use of evaluation?	None
GP-2. Are goals and purposes consistent with sound judicial principles, mission, and needs of the public?	P-1: Service Orientation U-1 Constructive Orientation
GP-3. Is Program structured and implemented so it does not impair the independence of the judiciary?	P-1: Service Orientation F-1 Practical Procedures
GP-4. Are additional (secondary) purposes recognized and considered appropriate by "governing" body?	U-2 Defined Uses
Administration and Support (AS)	
AS-1. Is responsibility for program development and implementation vested in highest court?	None
AS-2. Is day-to-day implementation monitored by broad-based group of individuals representing judges, lawyers, and non-lawyers familiar with the judicial system?	F-2 Political Viability
AS-3. Is program adequately funded and staffed?	F-3 Fiscal Viability
Performance Criteria (PC)	
PC-1. Are judges evaluated on integrity, including:	
PC-1a. Avoidance of impropriety?	A-1 Defined Role
PC-1b. Freedom from personal bias?	A-1 Defined Role
PC-1c. Decisions on issues not influenced by external pressures?	A-1 Defined Role
PC-1d. Impartiality?	A-1 Defined Role

ABA Guidelines, adapted checklist	Personnel Standards
PC-2. Are judges evaluated on knowledge, understanding, and execution of the law, including:	
PC-2a. Issuance of legally sound decisions?	A-1 Defined Role
PC-2b. Substantive, procedural, and evidentiary law?	A-1 Defined Role
PC-2c. Factual and legal issues before the court?	A-1 Defined Role
PC-2d. Application of judicial precedents, and other sources of authority?	A-1 Defined Role
PC-3. Are judges evaluated on communication skills, including:	
PC-3a. Clarity of bench rulings and other oral communications?	A-1 Defined Role
PC-3b. Quality of written opinions; clarity and logic?	A-1 Defined Role
PC-3c. Sensitivity to one's impact relating to demeanor, non-verbal communication?	A-1 Defined Role
PC-4. Are judges evaluated on judicial management skills, including:	
PC-4a. Preparation, attentiveness, and control over proceedings?	A-1 Defined Role
PC-4b. Devoting appropriate time to all pending matters?	A-1 Defined Role
PC-4c. Discharging administrative responsibilities diligently?	A-1 Defined Role
PC-4d. Management of calendar (e.g., number, age, and status of pending cases)?	A-1 Defined Role

ABA Guidelines, adapted checklist	Personnel Standards
PC-4e. Punctuality (i.e., prompt disposition of pending matters while maintaining rules of court)?	A-1 Defined Role
PC-5. Are judges evaluated on courtroom demeanor, including:	
PC-5a. Courtesy?	A-1 Defined Role
PC-5b. Willingness to permit parties to be heard?	A-1 Defined Role
PC-6. Are judges evaluated on service to profession and to public, including:	
PC-6a. Participation in judicial education programs?	P-1: Service Orientation
PC-6b. Assurance to the public that members of the judiciary serve to the best of their ability?	P-1: Service Orientation
PC-7. Are judges evaluated on effective working relationships with other judges, including:	
PC-7a. Exchange of ideas, opinions with other judges (when party of a multi-judge panel)?	None
PC-7b. Sound critiquing of colleagues' work?	None
PC-7c. Facilitating performance of other judges' administrative responsibilities?	None
Methodology	
M-1. Was Program developed systematically?	None
M-2. Is program execution sufficiently flexible, allowing improvements to be made?	None
M-3. Is Program periodically assessed?	A-8 Monitoring Evaluation Systems

ABA Guidelines, adapted checklist	Personnel Standards
M-4. Are appropriate, additional criteria for performance evaluation developed for use in jurisdictions (courts) that have unique characteristics and specific needs?	A-1 Defined Role A-2 Work Environment
M-5. Does the evaluation process include data collection, synthesis/analysis, and usage?	P-2 Formal Evaluation A-3 Documentation of Procedures
M-6. Are the analyses, evaluation timetable, and use of evaluation results appropriate for type of jurisdiction and extent of judges' experience?	None
M-7. Are methods for data collection and analysis developed with assistance with (measurement) experts to ensure quality?	None
M-8. Are reliable, valid and multiple sources of evaluation information employed?	A-4 Valid Measurement A-5 Reliable Measurement
M-9. Is information on performance based on "personal and current knowledge?"	P-3 Conflict of Interest U-3 Evaluator Credibility
M-10. Are provisions for confidentiality (of judges' ratings and respondents' identities) established?	P-4 Access to Personnel Evaluation Reports A-6 Systematic Data Control
Uses and dissemination	
UD-1. Is the dissemination of results consistent with Program's purpose?	U-4 Functional Reporting
UD-2. Is confidentiality of data and results maintained?	P-4 Access to Personnel Evaluation Reports A-6 Systematic Data Control
UD-3. Are results shared with individual participating judges as well as senior or administrative judges?	U-4 Functional Reporting U-5 Follow-Up and Impact
UD-4. Are unwarranted and potentially misleading analyses of individual judges avoided?	U-1 Constructive Orientation A-7 Bias Control
UD-5. If additional uses of evaluation results are required, are results provided to responsible parties without the promotion of a particular philosophy?	P-5 Interactions with Evaluatees U-5 Follow-Up and Impact
UD-6. If additional uses of evaluation results are required, are results provided to responsible parties after judges are afforded an opportunity to review and comment on the results?	P-5 Interactions with Evaluatees U-5 Follow-Up and Impact
UD-7. Does the use of performance evaluation exclude the use of results for purposes of discipline?	U-1 Constructive Orientation U-2 Defined Uses