State judicial performance evaluation (JPE) programs promise to help courts achieve a variety of central goals (e.g., more informed judicial selection, retention, and/or assignment decisions; improvements in judicial quality; greater transparency). However, recent criticisms leveled against these programs and supported by preliminary empirical evidence portray JPE surveys based on the popular ABA model as systematically biased against minority and women judges. Such claims invariably heighten the methodological scrutiny applied to all survey-based JPE programs that, in turn, will likely reveal a number of other shortcomings in existing JPE surveys. Most state JPE surveys do not reflect recent advances in the scientific understanding of survey design related to performance evaluation. States must remedy weaknesses in their JPE surveys if they wish to preserve the credibility of JPE programs in the public’s eye and within the court community. To provide states with some guidance in this effort, we review several fundamental shortcomings common to state and model JPE surveys in the U.S. and offer some concrete steps for improvement in key areas.

An Overview of Judicial Performance Evaluation in the States

The first official state-sponsored judicial performance evaluation program began in Alaska in 1976 as part of an effort to address concerns that the voting public lacked sufficient information to make educated decisions about judges in retention elections. Many other states followed suit: A 2004 national survey identified 21 states and territories with official JPE programs and 1 state with a pilot program. The Institute for the Advancement of the American Legal System (IAALS) website recognizes 18 states that presently have active JPE programs. The specific purposes of these programs vary by state: Results may be disseminated to judges to facilitate self-improvement, to the public to facilitate more informed voting decisions, and to judicial administration to facilitate more effective retention decisions and inform other administrative decision-making processes. Proponents of JPEs point to the potential of these programs to improve the quality of justice, to reinforce judicial independence, and to foster greater public

We thank Eric Heggestad, Margaret Stockdale, and two anonymous reviewers for their insights and valuable feedback. Correspondence concerning this article should be addressed to Jennifer K. Elek at the National Center for State Courts Research Division, 300 Newport Avenue, Williamsburg, VA, 23168 or jlek@ncsc.org.

4. Rottman & Strickland, State Court Organization 2004 (Washington, D.C.: U.S. Department of Justice, 2006). It is also worth noting that 9 additional states in this survey reported having JPE programs operated independently by their state bar associations.
Whether JPE programs reach their full potential depends on a variety of factors. The foundation of a strong JPE program rests perhaps most heavily on the quality of metrics used. Any state seeking to create or improve upon an existing JPE program faces a number of critical decisions about the substantive content and execution of the performance evaluation itself. Early in the development process, for example, those responsible for a state JPE program must choose which of the available performance measures and methods to employ. These may include but are not limited to caseload and workload statistics, courtroom observations, personal interviews, substantive reviews (e.g., of the judge’s decisions, opinions, or orders; of past disciplinary actions of the judge’s decisions, opinions, or orders; of past disciplinary actions), empirical research on data drawn from one unofficial JPE survey, for example, shows that women and members of minority groups receive more negative evaluations than their white male counterparts, legitimizing anecdotal claims made about gender and racial bias in the JPE survey process more generally. The quality of some state JPE programs has been questioned even by those involved in and generally supportive of the JPE process. For example, in a 2008 opinion survey of judges from a state with a long-running and respected JPE program, only 12.3% felt that the “validity and accuracy of survey responses were not a problem.” To stimulate discussion and possible corrective action in response to these concerns, this article offers detailed guidance on how states can develop a JPE survey instrument and process that minimizes bias and produces high-quality, meaningful evaluation data on judicial performance. First, we identify some significant concerns about the design of existing state and model JPE survey methods and explain why and how particular shortcomings in survey development and implementation can introduce bias into the evaluation process. We then offer some suggestions on how to implement best practices in performance measurement and survey design to address these issues.

The Quality of Existing Judicial Performance Evaluation Surveys

Although a layperson may view the development of a survey instrument as a relatively simple and straightforward process, in reality, seemingly trivial language, design, and implementation decisions can seriously compromise survey quality. These decisions can create coverage, sampling, non-response, and measurement errors that reduce the accuracy of results. The first three types of errors result when an unrepresentative sample of respondents is used—that is, when the experiences or opinions gathered from survey respondents do not accurately reflect those of the entire population of all potential respondents (e.g., a national poll of voter attitudes that uses a predominantly Republican sample of respondents). On the other hand, measurement error, or how much a measured value (e.g., results from a national poll of how much voters trust and confidence in the judiciary, among other benefits).

Of the available methods, nearly every state incorporates some form of evaluation survey. Undoubtedly, the popularity of the survey method is due in part to the fact that surveys can be used to gather a standard set of information from a large group of individuals in a relatively short amount of time. A well-designed survey is a powerful and efficient data-collection tool. However, poorly designed surveys can produce misleading or useless information.

Despite the many useful conceptual guides that have been proffered to advise states on JPE program development and to outline the many indicators of good judicial performance, little detailed technical guidance is publicly available to states on how to craft a good JPE survey instrument. In the absence of formal guidance on instrument development, some state committees appear to develop JPE survey instruments from scratch or simply replicate those in use elsewhere, often with limited input from scientific experts in survey methods or job performance evaluation. Without strong participation from such experts, a JPE survey is likely to fall prey to any number of common survey design mistakes that undermine the quality of survey data. Moreover, JPE surveys should be revised or redesigned to reflect important recent refinements to methodologies for survey-based job performance evaluation.

The time has come to review the quality of existing JPE survey methods. New evidence of problems with existing JPE surveys is starting to emerge. Empirical research on data drawn from one unofficial JPE survey, for example, shows that women and members of minority groups receive more negative evaluations than their white male counterparts, legitimizing anecdotal claims made about gender and racial bias in the JPE survey process more generally. The quality of some state JPE programs has been questioned even by those involved in and generally supportive of the JPE process. For example, in a 2008 opinion survey of judges from a state with a long-running and respected JPE program, only 12.3% felt that the “validity and accuracy of survey responses were not a problem.” To stimulate discussion and possible corrective action in response to these concerns, this article offers detailed guidance on how states can develop a JPE survey instrument and process that minimizes bias and produces high-quality, meaningful evaluation data on judicial performance. First, we identify some significant concerns about the design of existing state and model JPE survey methods and explain why and how particular shortcomings in survey development and implementation can introduce bias into the evaluation process. We then offer some suggestions on how to implement best practices in performance measurement and survey design to address these issues.

The Quality of Existing Judicial Performance Evaluation Surveys

Although a layperson may view the development of a survey instrument as a relatively simple and straightforward process, in reality, seemingly trivial language, design, and implementation decisions can seriously compromise survey quality. These decisions can create coverage, sampling, non-response, and measurement errors that reduce the accuracy of results. The first three types of errors result when an unrepresentative sample of respondents is used—that is, when the experiences or opinions gathered from survey respondents do not accurately reflect those of the entire population of all potential respondents (e.g., a national poll of voter attitudes that uses a predominantly Republican sample of respondents). On the other hand, measurement error, or how much a measured value (e.g., results from a national poll of how much voters trust and confidence in the judiciary, among other benefits).

Of the available methods, nearly every state incorporates some form of evaluation survey. Undoubtedly, the popularity of the survey method is due in part to the fact that surveys can be used to gather a standard set of information from a large group of individuals in a relatively short amount of time. A well-designed survey is a powerful and efficient data-collection tool. However, poorly designed surveys can produce misleading or useless information.

Despite the many useful conceptual guides that have been proffered to advise states on JPE program development and to outline the many indicators of good judicial performance, little detailed technical guidance is publicly available to states on how to craft a good JPE survey instrument. In the absence of formal guidance on instrument development, some state committees appear to develop JPE survey instruments from scratch or simply replicate those in use elsewhere, often with limited input from scientific experts in survey methods or job performance evaluation. Without strong participation from such experts, a JPE survey is likely to fall prey to any number of common survey design mistakes that undermine the quality of survey data. Moreover, JPE surveys should be revised or redesigned to reflect important recent refinements to methodologies for survey-based job performance evaluation.

The time has come to review the quality of existing JPE survey methods. New evidence of problems with existing JPE surveys is starting to emerge. Empirical research on data drawn from one unofficial JPE survey, for example, shows that women and members of minority groups receive more negative evaluations than their white male counterparts, legitimizing anecdotal claims made about gender and racial bias in the JPE survey process more generally. The quality of some state JPE programs has been questioned even by those involved in and generally supportive of the JPE process. For example, in a 2008 opinion survey of judges from a state with a long-running and respected JPE program, only 12.3% felt that the “validity and accuracy of survey responses were not a problem.” To stimulate discussion and possible corrective action in response to these concerns, this article offers detailed guidance on how states can develop a JPE survey instrument and process that minimizes bias and produces high-quality, meaningful evaluation data on judicial performance. First, we identify some significant concerns about the design of existing state and model JPE survey methods and explain why and how particular shortcomings in survey development and implementation can introduce bias into the evaluation process. We then offer some suggestions on how to implement best practices in performance measurement and survey design to address these issues.

The Quality of Existing Judicial Performance Evaluation Surveys

Although a layperson may view the development of a survey instrument as a relatively simple and straightforward process, in reality, seemingly trivial language, design, and implementation decisions can seriously compromise survey quality. These decisions can create coverage, sampling, non-response, and measurement errors that reduce the accuracy of results. The first three types of errors result when an unrepresentative sample of respondents is used—that is, when the experiences or opinions gathered from survey respondents do not accurately reflect those of the entire population of all potential respondents (e.g., a national poll of voter attitudes that uses a predominantly Republican sample of respondents). On the other hand, measurement error, or how much a measured value (e.g., results from a national poll of how much voters
like a presidential candidate) differs from its actual value (e.g., how much voters actually like the candidate), is exacerbated by poor instrument construction. We focus primarily on the latter type of error, but will touch on elements from all four.

Decisions that mold the content of the survey instrument frame the context for evaluation and establish the process by which respondents formulate their judgments. By clearly defining the target behavior under review and the standards for use in evaluating that target behavior, instrument developers can minimize measurement error. Without clear definitions that instruct attorneys and others to evaluate the judge on his or her actual behavior, however, people will tend to base judgments on expectations (e.g., assumptions derived from rumors or personal reputation), stereotypes (e.g., about judges in general) and other heuristics that may produce systematic bias.\(^{16}\) As a result, JPE surveys may then be susceptible to a wide range of well-documented response biases that are common with the survey method.\(^{17}\)

Although research on the efficacy of state JPE survey instruments is sparse,\(^{16}\) systematic gender and racial disparities have been found in JPE survey results in some states.\(^{19}\) These findings are reinforced by a larger body of scientific research which shows that people make judgments of others that often are biased (intended or not) by cultural stereotypes in a number of ways. For example, a job may be seen as requiring specific skills or traits that are associated with a particular stereotype about a certain social group.\(^{20}\) When judgments require respondents to ascribe personality characteristics to an individual or develop higher-order attributions about the individual’s personality traits or abilities, these judgments may be informed by such stereotypes.\(^{21}\) One study illustrates this phenomenon well: Participants tended to explain a male’s successful performance on a “masculine” task as indicative of his ability, but tended to explain a female’s successful performance on the same task as a result of pure luck.\(^{22}\) Moreover, stereotypes may subtly alter the standards used in evaluation if the provided standards are poorly defined. In one empirical study, people asked to evaluate a female and a male candidate for a “masculine” police chief position reported that whichever qualifications the male applicant possessed that the female applicant did not (street smarts, formal education) were more important for the job.\(^{23}\) Thus, biases may stem not only from any explicit prejudices a person may have, but also from implicit associations derived from cultural stereotypes and other heuristics that simplify information processing for the respondent.

With these concerns in mind, we examined 18 current or recently used state JPE surveys and four “model” surveys (see Appendix A) for their use of basic best practices in survey design. Recall that state JPE programs may serve a range of purposes from informing judicial retention and assignment decisions to informing judges for educational purposes and results may be strictly confidential or publicly published and disseminated. The selection of surveys we reviewed represents this diversity. Moreover, state JPE programs most commonly target attorney respondents; most survey instruments of other respondent groups tend to be either similar in structure to the attorney survey or substantially less comprehensive. For the purpose of this review, we focused on the attorney versions of JPE surveys. Despite some variation in content, these surveys tended to share some fundamental design flaws that could produce low-quality data. The most common and important of these flaws are discussed below.

**Survey content issues.** A number of state and model JPE surveys exhibit problems in how the question stems were constructed and with the types of standard response options provided:

- **Poor item construction.** In the
JPE surveys we reviewed, poorly constructed question stems were a common problem. Some of the most pervasive issues included (a) the double-barreled problem, (b) language that is too abstract or vague, and (c) language that does not match the provided response options. Explanations and examples of each of these issues follow.

Double-barreled items. One problem with many existing JPE surveys is that they are littered with double-barreled items. A double-barreled item is a compound question that presents two or more qualitatively different aspects of judicial behavior for evaluation but that only allows the respondent to supply a single rating response. These types of problematic items can often be identified by use of the conjunction “and.” For example, a few of the double-barreled items we found in the set of existing state and model JPE surveys we reviewed include:

- “The judge listens with patience and attentiveness”
- “Uses common sense and is resourceful in resolving problems that arise during proceedings”
- “Is effective in initiating and managing settlement efforts/conferences”
- “Providing written communications that are clear, thorough, and well reasoned”
- “Is able to maintain proper order, decorum, and control in the courtroom”
- “Judge acts with patience, dignity, and courtesy”
- “Oral and written decisions and orders are clear and well reasoned”

Surveys that contain double-barreled items are poor evaluation instruments because they invariably produce inaccurate results. For each item, each respondent must subjectively decide which embedded question to answer. Consider the item above that refers to the judge’s degree of patience, dignity, or courtesy (a triple-barreled item). How is a respondent to rate the judge if the judge acts with dignity and courtesy but is not very patient? Perhaps the respondent will rate the judge on one of the three elements (patience or dignity or courtesy), or average across the three, or take a completely different tack. The respondent does not have a clear idea of what question is truly being asked in the survey, and the judge, the state JPE committee, and the public will not know what respondent ratings actually mean about the judge’s performance on double-barreled items. Moreover, those judges who do not receive perfect ratings will also not be able to discern from the item exactly which behavior(s) to target for improvement. By failing to clearly communicate unambiguous feedback about which judicial behaviors should be changed, many JPE programs designed with a purpose of improving performance quality do not efficiently fulfill this goal. The elimination of confusing compound items is one easy step toward a clearer, more functional, and more meaningful survey-based JPE program.

Language that is too abstract or vague. Survey items should be articulated clearly using specific, concrete language. Many state JPEs are plagued with language that is too abstract or vague. As discussed above, respondents must impose their own subjective definitions onto abstract or ill-defined items, making subsequent ratings for those items difficult to interpret. Some examples of vague question stems include:

- “Judge’s charge to the jury/juries”
- “Gives proceedings a sense of dignity”
- “Demonstrates appropriate demeanor on the bench”
- “Effective as an administrator”
- “The judge promotes public confidence in the judiciary”
- “Behaves in a manner that encourages respect for courts”
- “Demonstration of appropriate compassion”

What exactly is the respondent rating about “the judge’s charge to the jury”? What does “the judge promotes public confidence in the judiciary” or “gives proceedings a sense of dignity” mean in terms of actual, observable behavior? What is the precise definition of “appropriate compassion” or “appropriate demeanor” for a judge? In each of these cases, it is impossible to know what the judge is actually doing to earn the performance rating that s/he receives. These items fail to adequately describe the types of behavior under evaluation and performance standards necessary to rate that behavior in a meaningful way.

An instrument designed to measure “performance” should focus on observable behavior. Some items from existing instruments, however, employ abstract language that implicitly requires respondents to make assumptions about or attribute general characteristics to the judge in order to produce a rating response. Instead of focusing on how the judge actually behaves in court, some of these items also require respondents to conjecture about what and how the judge thinks. For example, consider the following question stems from existing state JPE surveys:

- “Is willing to reconsider error in fact or law”
- “Willingness to make difficult or unpopular decisions”
- “Skills in effecting compromise”
- “Patience”
- “Keeps an open mind and considering all relevant issues in making decisions”

Other problems with these items aside, most respondents probably do not have accurate, unbiased insight about what a judge is willing or not willing to do. Instead, they must use other information to develop a judgment about the likelihood that the judge thinks in a particular way. Rephrasing these items to capture this observable information would produce cleaner, more reliable data and ultimately better guidance on which behaviors should be reinforced and which behaviors should be changed.

Additionally, items that require respondents to form higher-order attributions about the judge’s overall skill, ability, knowledge, or personality (such as those that ask respondents to evaluate the judge’s “patience” or “skills in effecting compromise”) can elicit ratings that are systematically distorted. In the process of forming these higher-order judgments of others, people—particularly those
with fewer opportunities to interact with the judge—often rely on more than just observed behavior. They tend to interpret and categorize others using information derived from social stereotypes and other heuristics.\textsuperscript{24} These stereotypes and other assumptions can color judgment and bias evaluations of performance.\textsuperscript{25}

**Poor rating scales.** In the 18 state and four model JPE surveys we reviewed, a poorly constructed rating scale emerged as one of the most pervasive problems. Poor rating scales typically resulted from conflating two or more evaluation constructs, failing to provide adequate definitions for scale points, or a combination of both.

When a rating scale conflates two or more constructs, respondents may experience difficulty distinguishing between response options on a consistent basis. One example of a conflated response scale that is used by some state JPE programs is the academic grading scale (A-D, F), incongruously defined in part by some level of acceptability and some degree of performance quality, such as:

- A = Excellent
- B = Very Good
- C = Acceptable
- D = Poor
- F = Unacceptable

A respondent forced to answer using this rating scale may have trouble determining whether a poor performance in a particular area should be scored as “poor” or as “unacceptable” if s/he believes a poor performance is unacceptable. Or perhaps the respondent may view only an excellent performance as acceptable, but this subjective working definition may change for a different performance area. The differences between available response options are unclear and without further elaboration rely on each individual respondent’s subjective interpretation of how the scale should be applied in each case, to the detriment of instrument reliability.

As vague survey items contribute to measurement error, so do vague or inadequately defined rating scale options. That is, some scales may not conflate different constructs as in the example above, but they may not provide a clear description of how the rating scale options should be applied to observed behavior either. For example, Kourlis, Gagel, Singer, and colleagues\textsuperscript{26} argue that Alaska’s JPE survey rating scale does not provide respondents with a definition for the “minimum standards of performance for this court:”

1. **Poor:** Seldom meets minimum standards of performance for this court.
2. **Deficient:** Does not always meet minimum standards of performance for this court.
3. **Acceptable:** Meets minimum standards of performance for this court.
4. **Good:** Often exceeds minimum standards of performance for this court.

Without a clear understanding of what the “standards of performance for this court” are and whether or how these standards may differ between individual courts across the state, the provided scale point definitions fail to clarify the meaning of each response option or denote when such an option would be appropriate.

Of course, some evaluation surveys do not define the provided rating scale at all. In particular, two respected state pioneers of JPE used survey instruments that offered respondents only a grading scale (A-D, F). These surveys did not supply any definition or instructions for how the grading scale should be applied by respondents in the evaluation of judicial performance, or what type of performance would earn a judge an “A” rating.

When presented with conflated and/or vague rating scales, respondents must draw on available extraneous information to simplify and clarify the otherwise challenging performance evaluation task. What information respondents draw on to clarify this task varies across individuals and can even change for the individual respondent as he or she progresses through the evaluation survey. Thus respondents are more likely to respond in a manner that is systematically biased in favor of overall leniency, middle-of-the-road options, or particular social groups, to the detriment of data quality. Greater subjectivity in how response options are interpreted and used produces greater variability in the survey data; thus poorly designed JPE surveys can be highly unreliable sources of information about performance quality.

Items incongruous with the provided rating scale and other errors.

---


\textsuperscript{26} Kourlis et al., *Shared Expectations*, supra n. 7.
Some JPE surveys list question stems that do not match some or all of the supplied response options. A mismatch occurs when an item is phrased in such a way that not every response along the provided rating scale clearly or logically follows. For example, one question stem we found asked respondents to assess the judge on “absence of coercion or threat” using a 5-point scale of “excellent,” “good,” “adequate,” “less than adequate,” and “poor.” It is challenging to match rating scale items with performance on this item. Other problems with this item aside, a more appropriate item-response scale pairing would include the question stem like “uses coercion or threats” and yes/no response options or a response scale to gauge frequency of occurrence. Similarly, another state JPE survey asks respondents to rate judges on “judicial temperament” along a 5-point scale of “poor,” “deficient,” “acceptable,” “good,” and “excellent.” However, some of these options are non-sequiturs (e.g., how a person can have a deficient temperament is unclear). Although this typically occurs in conjunction with a poorly crafted rating scale, a mismatch may occur even if the provided rating scale is clear and unidimensional.

Incongruous item-scale pairs can force the respondent to make subjective decisions about how to interpret the question and how to apply each response option. Like the other survey construction errors, this problem can hurt the actual validity of the survey by reducing response accuracy. It may also hurt the face validity of the survey (i.e., layperson perceptions of survey quality) by leading people to conclude that the instrument is not a good one or does not work as advertised.

Finally, several surveys we reviewed would benefit from more thorough copy-editing before re-use. We found a substantial number of grammatical and typographical errors that may serve as another cue of poor survey quality to respondents, judges, and the public. If people perceive the JPE survey to be ineffective or of poor quality, they will not support the JPE program.

**Procedural issues.** Some JPE survey programs we reviewed incorp- orated procedures that could have an adverse effect on the quality of data obtained from the survey. These procedures could discourage honest responses, encourage biased responses, or encourage participants to respond based on biased sources of information. In addition, some JPE programs suffer from coverage, sampling, or nonresponse error; all of which can produce biased survey results.

**Formats that deter negative feedback.** Some surveys require respondents to answer additional questions only when they supply negative feedback about the judge. Two state JPE surveys we reviewed required respondents to explain their reasoning any time they used one of the lowest two rating options to describe the judge’s performance. This practice can imply that negative feedback should be used infrequently and that neutral or positive feedback is expected. Some respondents may wish to avoid the time burden or greater effort required to supply a clear argument for a negative rating choice. As a result, these respondents may be less likely to provide negative feedback (artificially inflating the ratings provided) or even less likely to complete the survey if the costs associated with honest participation outweigh the perceived benefits.

**Formats that encourage evaluation based on biased sources of information.** Some surveys are designed in a way that explicitly permits or encourages the use of unreliable or biased sources of information in respondent evaluations of a judge. Two of these formats and the problems associated with them are discussed below.

**Hearsay or personal contact.** In some jurisdictions, attorneys, court staff, and others can evaluate a judge’s performance even if they have not had any direct, professional contact with the judge. For example, one state JPE survey we reviewed allowed respondents to rate a judge even if they know the judge only by professional reputation or only through personal (i.e., non-professional) contact. Although this information is reported separately in the JPE report for this state, Kourlis, Gagel, Singer, and colleagues27 criticize this use of hearsay and non-judicial behavior in an evaluation of a judge’s workplace performance. Certainly, a respondent who has no direct experience with a judge in the professional workplace is ill-equipped to evaluate his or her judicial performance.

**Individual vs. grouped evaluations of judges.** Some JPE surveys ask respondents to evaluate multiple judges simultaneously on the same form. When making sequential judgments like this, respondents tend to engage in comparative thinking between the individuals under evaluation.28 These comparison-based assimilation and contrast effects may then artificially increase or decrease performance ratings of certain judges. For example, a respondent may give a particularly excellent judge the highest possible rating, and then inadvertently rate the next judge—a very good but not excellent judge—as somewhat higher than he or she might otherwise respond because of salient similarities with the preceding judge (assimilation effect). Or, a respondent may give a particularly terrible judge the lowest possible rating, and then rate the next judge—an average performer—as excellent in comparison (contrast effect). Thus, some judges may receive artificially lower or higher scores

27. Kourlis et al., SHARED EXPECTATIONS, supra n. 7.

simply because of the order in which they are rated by respondents in the same sitting. Moreover, a survey in which respondents self-select which judges they wish to evaluate has poor quality control over the sample of respondents who choose to evaluate any particular judge. Respondents may evaluate more familiar judges based on direct experiences, but also evaluate less familiar judges using information derived from their reputations or from heuristics about other differentiating personal characteristics, such as gender or racial stereotypes. Evaluation surveys of multiple judges, therefore, may produce less accurate results than surveys in which respondents are asked to evaluate one particular judge independently of the others.

Unrepresentative data. The quality of a survey depends on the representativeness of resulting data to the opinions of the target population (i.e., the amount of survey error). Although survey methods in the real world are often informed by practical considerations such as total population size, convenience samples, availability of contact information, and program costs, such considerations should be evaluated carefully given their inherent trade-offs with data quality. Undoubtedly, each state or jurisdiction struggles with a different array of limitations or challenges in a data collection process that requires careful consideration to ensure that the JPE survey is implemented in a manner that reflects the best possible compromise between survey best practices and the practical realities of applied data collection. However, those overseeing state JPE programs that rely on survey-based methods should recognize that if survey data are collected from a group of respondents who differ in some systematic way from others in the target population who were not given the opportunity to complete the survey or who opted not to participate in the survey, the results may lead to erroneous conclusions about the opinions of the target population. Some state JPE programs may suffer from any or all three of the following sources of survey error, each of which may contribute to the problem of unrepresentative data.

Coverage error. Coverage error can occur “when not all members of the population have a known, nonzero chance of being included in the sample for the survey and when those who are excluded are different from those who are included on measures of interest.”29 That is, the survey may be based on data from a subset of individuals that is not representative of the larger population of potential respondents because certain types of individuals are systematically excluded from the surveying process. These individuals may be inadvertently excluded because the survey mode (e.g., web-based surveys) makes the survey inaccessible to certain types of individuals (e.g., people who prefer not to use email or the internet, or who do not have reliable internet access) whose opinions may differ from those who do have access to the survey. Alternatively, the source of potential respondents from which a sample is drawn (e.g., outdated respondent lists that exclude new attorneys; lists that contain only individuals who voluntarily registered with a particular organization or who allowed their contact information to be publicly shared) may categorically exclude certain types of individuals, producing a biased sample. In either case, the resulting sample may not be representative of the entire population to which the researchers intend to generalize the survey results. Some JPE programs, particularly in states without a statewide database that documents attorney appearances before each judge, may struggle with coverage error when attempting to create viable lists of respondents from which to sample. They may then be forced to adopt more costly and labor-intensive methods, such as assembling lists piecemeal from various impoverished sources to create a more complete directory of the eligible population, or to accept the problems associated with existing convenience samples.

Sampling error. In some states, the entire population of eligible respondents may be invited to complete a JPE survey, but typically, efforts are focused on a smaller subset of individuals because surveying the entire population is too costly. Sampling error “results from surveying only some rather than all members of the population and exists as a part of all sample surveys.”30 That is, results of any survey based on a smaller sample of respondents will necessarily be an estimate of the opinions of the larger population that, as with any estimate, contains some error. Thus surveys based on very small effective sample sizes may produce results that are not indicative of the opinions of the entire eligible population. Perhaps due to the logistical problem of coverage limitations (discussed above), some programs invited small numbers of attorneys or other respondents to participate in each judge’s JPE survey. In one pilot study, fewer than 10

29. Dillman et al., supra n. 11, at 17.
30. Id.
surveys were delivered to potential attorney respondents for nearly half of all evaluated judges; the majority of these judges had surveys delivered to five potential respondents or less. If random samples are of sufficient size, however, they can produce estimates of the opinions of the larger population within an acceptable margin of error.

Non-response error. Even if the JPE program committee elects to survey the entire population of potential respondents or devises a good representative random sample from unbiased lists of potential respondents, certain types of individuals may tend to accept the invitation to participate, whereas certain other types of individuals may tend to opt out. That is, some subgroups of eligible participants may be more likely to self-select into completing the survey than others and their opinions may thus be overrepresented in the data. One state issued their JPE survey to all attorneys registered with the state bar association and allowed each attorney to choose whether or not to complete an evaluation survey for each judge, with no evident attempt to verify subsequent data quality. Compounding the problem, JPE surveys have notoriously poor response rates. Many modern JPE programs fail to achieve the often-cited acceptable response rate of 50% and some are reporting seriously poor response rates. Many program committees should scrutinize the rating scale or scales currently in use, as this was a significant non-response error, limiting the extent to which researchers and others can generalize from the sample to the larger population. Although low response rates may not necessarily always result in non-response error, a well-designed survey process should take precautions to minimize non-response rates in order to reduce the possibility of this type of error to adversely affecting data quality.

Recommendations for Improvement

States can improve the accuracy of existing JPE surveys in a number of ways:

**Use simple, concrete items that describe observable behaviors.** Instrument developers can follow several useful guidelines when preparing survey items for use. First, they should avoid creating or adopting survey items that are too complex. Items that contain the conjunction “and” can be simplified by replacing similar terminology with a single catch-all phrase or by separating distinct terms into different question stems. This practice should reduce the incidence of problematic double-barreled items.

In addition, instrument developers should ensure that all provided response options follow logically from the question stems. For example, every option along a frequency rating scale of “Never,” “Rarely,” “Sometimes,” “Often,” and “Always” would make sense as a response to the evaluation item, “The judge listened carefully during the court proceeding.” If a single, standard set of response options does not apply to every question, instrument designers should consider either rephrasing questions in a way that all response options apply, reworking the rating scale so that the options match the existing questions, or creating separate rating scales to match individual questions or sets of questions.

Most importantly, selected items should contain concrete language that describes behaviors that respondents could reasonably observe in their direct experiences with the judge. Question stems that focus on specific, observable behavior (e.g., “The judge started courtroom proceedings on time”), rather than inferred attributes (e.g., “The judge is patient”) and/or vague, abstract concepts (e.g., “The judge gives proceedings a sense of dignity”), will tend to produce more accurate ratings that are less influenced by stereotypes, expectations, and other heuristics.

Some innovative techniques demonstrate great promise in further reducing response biases and stereotype bias in surveys. One particular strategy incorporates a structured free-recall (SFR) task that forces evaluators to actively recall specific memories about a person’s past behavior prior to completing the evaluation of his or her performance. In the SFR task, respondents are prompted to recall and then list specific positive and negative behaviors they observed from the evaluated person. This intervention facilitates evaluator recall of observed behavior, reduces evaluator reliance on heuristics when subsequently formulating judgments, and increases overall response accuracy.

**Improve rating scales.** State JPE program committees should also scrutinize the rating scale or scales currently in use, as this was a significant problem in many of the survey instruments we reviewed. Poor rating scales should be modified or replaced entirely.

A good rating scale: *Is balanced.* If a bipolar construct is

33. E.g., id.
39. Baltes, Bauer & Frensch, Does A Structured Free Recall Intervention Reduce the Effect of Stereotypes on Performance Ratings and by What Cognitive Mechanism?, 92 J. Applied Psychol. 151 (2007); Baltes & Parker, Clarifying the Role of Memory in the Performance Cue Effect, supra n. 25; Baltes & Parker, Reducing the Effects of Performance Expectations on Behavioral Ratings, supra n. 25; Bauer & Baltes, supra n. 38.
used (i.e., one that measures an attribute from one end of the value spectrum to the other, such as extremely negative to extremely positive), the scale should provide respondents with an equal number of positive scale points as negative scale points. A balanced 7-point scale is typically most effective for bipolar scale types. Use of unbalanced response scales (i.e., response scales that contain more positive scale points than negative, or vice versa) can increase the likelihood of eliciting a response consistent with whichever type of response is more heavily represented in the rating scale. For example, the following unbalanced rating scale used by one state JPE survey contains only one unfavorable response option: “Excellent,” “Very Good,” “Satisfactory,” “Fair,” and “Unsatisfactory.” Instrument developers should thus revise such unbalanced scales to create a more balanced set of response options. One balanced 5-point revision of the above example is: “Very Good,” “Satisfactory,” “Fair,” “Poor,” and “Very Poor.”

If instrument developers can reasonably expect that only half of such a bipolar rating scale will be used by respondents and greater precision for that half of the rating scale is needed, a unipolar scale may be more appropriate. A unipolar scale captures the presence or absence of an attribute (e.g., from “not at all” to “extremely favorable”). A 5-point scale is typically most effective for unipolar scale types. One example of a unipolar scale is the following frequency scale, which ranges from never to always: “Never,” “Rarely,” “Sometimes,” “Usually,” and “Always.” Note that all scale points (i.e., response options) are meaningfully labeled with a descriptive adjective or other text, as fully labeled scales produce better quality data.

Is proportional. The qualitative meaning of the category labels or adjectives that represent the rating scale points should increase in increments that are proportional to one another. When selecting labels for each point on the rating scale, the conceptual distance between adjacent descriptive labels should be equivalent. For example, on a hypothetical 4-point scale of Poor / Fair / OK / Excellent, the conceptual distance (i.e., difference in qualitative meaning) between “poor” and “fair” is much larger than the conceptual distance between “fair” and “OK.”

The survey instrument should also present the scale using spacing that is visually proportional between each scale point along a single row. Spacing between response options serves as a visual cue regarding the conceptual distance between intervals on a scale, and greater spacing between some options and not others can bias responses towards whichever side of the scale occupies a greater proportion of visual space. In addition, presenting rating scale response options in multiple rows and columns rather than in a single row can unduly bias survey results. However, any response options that are not part of the linear rating scale (e.g., a “Not Applicable” option) should be visually distinguished from the scale in some way. Without additional spacing or a different format to set the N/A response option apart from the rating scale, the visual midpoint of the set of response options no longer represents the conceptual midpoint of the rating scale. The N/A response option may therefore be perceived as part of the rating continuum by a casual respondent, biasing results.

Survey items should contain concrete language that describes behaviors respondents observe in their interactions with judges. Measures a single construct. A well-crafted rating scale should provide scale points that indicate degrees of a single construct (e.g., performance quality) instead of one that conflates different constructs (e.g., a scale with options that alternate between degrees of performance quality and degrees of satisfaction with performance). When survey instrument developers create a rating scale by picking scale point labels from different measurement constructs, the resulting survey forces respondents to select a single response option from an array of options that are not necessarily mutually exclusive. If instrument developers want to capture multiple constructs (e.g., if they want an assessment of the quality of the judge’s performance as well as an assessment of whether or not that performance meets the respondent’s minimum personal standards), they should consider asking separate questions that employ different rating scales.

Keeps qualitatively distinct response options separate. Some surveys offer options that merge qualitatively different responses. For example, some

44. For guidance, see Krosnick & Fabrigar, Designing Good Questionnaires: Insights From Psychology (New York, NY: Oxford University Press, 1997).
45. Id.
47. Dillman et al., supra n. 11.
48. See id., for further discussion.
surveys may offer a single checkbox option for respondents to indicate if they are undecided or if the question does not apply to them. However, respondents for whom the question applies but who are nonetheless undecided about how to rate the judge can be very different from respondents for whom the question does not apply. This distinction can be meaningful, and it is generally best to err on the side of caution, gathering more detailed information when possible. After data is collected, data analysts can always merge two separate response options, but they can never disentangle different groups who used a single, conflated response option.

**Standardize the amount of effort required from respondents, regardless of their feedback.** By requiring survey participants to elaborate on their responses only when they provide negative or extreme ratings, instrument developers may inadvertently discourage some respondents from providing such feedback. The extra work involved in elaborating upon responses may result in fewer of these types of performance ratings, even if the ratings are warranted. Other respondents may feel that their written explanations or commentary jeopardize their anonymity in the survey. Respondents may end up selecting more moderate ratings than they would otherwise offer. If additional comments are permitted in the survey, all respondents should have an equal opportunity to clarify their ratings in order to avoid deterring some from offering particular types of feedback.

**Select respondents who have recent experience with the judge, and prompt them to use more reliable sources of information.** The JPE program should survey respondents with recent and preferably regular experience working with the judge. Respondents are more likely to accurately recall recent experiences, as memory degrades over time. If eligible participants can be identified and sampled in a representative manner (see Gather data from respondents who adequately represent the target population, below), the result will be more reliable, meaningful feedback than if the respondents have infrequent or minimal opportunities to observe and interact with the judge undergoing evaluation.

A JPE survey instrument should also be structured in a way that encourages respondents to base their judgments on the most reliable sources of information. First, the survey should contain specific instructions for respondents not to base evaluations on reputation or personal or social contact, but instead on professional behavior directly observed in the workplace environment. The goal is to assess observed performance, not reputation or hearsay. Secondly, a single evaluation should survey a respondent about one individual judge rather than the array of all judges undergoing evaluation. Separate evaluations of each judge should allow for more accurate results than surveys that ask respondents to evaluate an array of judges simultaneously.49

**Conduct a careful internal test of the survey.** The survey development team should carefully review the final draft of the survey to ensure that all grammatical and typographical errors are resolved, all instructions are clear and concise, all items and response options make logical sense, and the mechanics of the survey operate as intended.

**Conduct a pilot study.** Once a survey instrument has been carefully crafted, reviewed, and tested internally, instrument developers should conduct a pilot study to ensure that the JPE survey meets statistical standards for instrument reliability and validity prior to an official statewide launch.50 Moreover, a pilot study provides instrument developers with an opportunity to identify sources of confusion in the survey. Conducting cognitive interviews52 with a handful of eligible respondents may help uncover these problem areas and may also produce useful suggestions for improvement.

**Gather data from respondents who adequately represent the target population.** For the pilot and production survey studies, states should gather response data from a representative sample of participants.52 The identified sample should be large enough to permit the expected statistical analyses and should not be drawn from sources that systematically omit certain types of otherwise eligible respondents from participation. Generally, larger samples produce more precise estimates than smaller samples, but this precision does not increase linearly; the benefits afforded by increases in sample size eventually level off to achieve only marginal, if any, additional gains in accuracy. The point at which the costs outweigh the benefits depends on the size of the target population and the nature of the survey.


---

52. See Dillman et al, supra n. 11, for detailed guidance on crafting a representative sample and on minimizing coverage, sampling, and nonresponse error.
53. Id.
individual (rather than bulk) emails or contact letters with personalized salutations\(^54\) and multiple follow-up reminders in addition to the initial invitation to participate,\(^55\) that vary depending on the selected method for administering the survey. For example, how mailed surveys are assembled in the envelope may influence nonresponse, whereas login strategies may play a role in both response rate and quality of responses to web-based surveys.\(^56\) These and other implementation decisions can exacerbate or reduce problems with sample representativeness.

**Moving Forward**

Our primary goal is to raise awareness of the need for better judicial performance evaluation surveys. Although some state JPE programs employ surveys with greater methodological rigor than others, it is clear that the quality of existing state JPE surveys as a whole can be greatly improved. We have identified an array of fundamental challenges faced by state JPE survey programs and provided general technical guidance for improvement, but some states may require more context-specific technical guidance to optimize the quality of their survey-based JPE programs. Additional steps may be advisable. Once survey methods conform to best practices, others may wish to focus more on implementation concerns and less on survey construction issues. In addition, more states should consider implementing non-survey approaches to judicial performance evaluation measurement; these approaches should receive the same level of empirical scrutiny as the survey-based method. Each individual measure used in JPE programs—from survey measures to individual caseload and workload statistics to the review of a sample of written opinions—is an incomplete picture of judicial performance. When surveys, interviews, more objective caseload data, and other information about a judge’s work performance are used as part of a multi-method evaluation program, the strengths of each individual measure can theoretically compensate for the shortcomings of others. Future research efforts should prioritize the development of efficient multi-method evaluation programs for which quality surveys are but one component.

As states integrate scientific best practices into their JPE programs, they will gain more powerful, more persuasive, more informative, and more useful systems of judicial performance evaluation. For those concerned about the integrity of state JPE programs and the viability of these programs in the future, these are strong motivations to take corrective action now.

---

\(^54\) Barron & Yechiam, Private E-Mail Requests and the Diffusion of Responsibility, 18 COMPUTERS IN HUM. BEHAV. 507 (2002); Joinson & Reips, Personalized Salutation, Power of Sender, and Response Rates to Web-Based Surveys, 23 COMPUTERS IN HUM. BEHAV. 1372 (2007).

\(^55\) Cook, Heath & Thompson, A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys, 60 EDUC. & PSYCHOL. MEASUREMENT 821 (2000).

\(^56\) See Dillman et al., supra n. 11, at 263-266, 209-292.

---

**Jennifer K. Elek, Ph.D.**

is a research associate at the National Center for State Courts. Her research activities involve such topics as judicial performance evaluation, gender and racial fairness, offender risk and needs assessment, and problem-solving courts.

**David B. Rottman, Ph.D.**

is a principal researcher at the National Center for State Courts. His research interests include public opinion on the courts, innovations in judicial performance evaluation, and the role of procedural fairness principles in court reform.

**Brian L. Cutler, Ph.D.**

is a professor at the University of Ontario Institute of Technology. He has served as Editor of *Law and Human Behavior* and President of the American Psychology-Law Society.

---

**AJS Featured in The Giving Library**

The Giving Library, a project of the Laura & John Arnold Foundation, offers philanthropists an innovative way to enhance their strategic charitable giving. Donors can explore an online archive of video interviews to locate, study, compare and engage with hundreds of nonprofit organizations across the country. AJS is proud to be featured in this compelling new medium, which will help to increase the Society’s visibility among prospective donors nationwide.

To view the AJS videos, visit www.givinglibrary.org and search for “American Judicature Society.”

---