NATIONAL QUALITY FORUM

Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties

January 2011

NATIONAL QUALITY FORUM

Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties

CONTENTS

INTRODUCTION

The National Quality Forum (NQF) relies on four criteria for evaluating the suitability of quality measures for endorsement as voluntary consensus standards: *Importance to Measure and Report, Scientific Acceptability of Measure Properties*, *Usability*, and *Feasibility*. The second criterion, *Scientific Acceptability of Measure Properties*, is an important aspect of the successful use of publicly reported measures to improve performance. Scientific acceptability of measure properties refers to the reliability and validity of measures. The use of measures that are unreliable or invalid undermines confidence in measures among providers and consumers of healthcare. The goal of this document is to provide recommendations on what constitutes scientific acceptability of measure properties to assist participants in the measure evaluation process, including steering committee and technical advisory panel members, as well as measure developers. Guidance on scientific acceptability will facilitate a shared understanding of this complex and highly specialized subject.

Both empirical evidence and expert judgment play a role in measure evaluation. However, judgment can be best applied when those evaluating a measure have a thorough understanding of the evidence of scientific acceptability that does or does not exist. Evidence that a clearly specified measure produces credible results on performance comes from the basic measurement principles of reliability and validity. Although reliability and validity have always been included in NQF evaluation criteria, the criteria have not included specific guidance on 1) the scope of testing, 2) what tests of reliability and validity could be performed, and 3) how to weigh the results of this testing.

Task Force Charge

The NQF Task Force on Measure Testing was asked to address the following tasks:

- Identify the type of testing for scientific acceptability that should be conducted for various types of measures and data sources, and determine whether there are any acceptable alternatives to formal testing.

- Identify the type of testing that should be required prior to endorsement of measures specified for electronic health records (EHRs)—both measures originally developed using other data sources besides the EHR and new measures developed specifically for the EHR.
- Develop guidance for measure stewards/developers and NQF steering committees and technical advisory panels on adequate measure testing, interpretation of results, and information about testing that should be provided in the measure submission.
- Make recommendations for potential enhancements to the evaluation criteria.

BACKGROUND

NQF endorses quality measures intended for use in quality improvement as well as public reporting. Measure scores are used to make decisions about selecting and rewarding healthcare providers (e.g., by consumers and purchasers) and to identify opportunities for quality improvement (e.g., by providers). The level of confidence that one has in the conclusions about quality based on the measure scores is a function of the reliability and validity of measurement.

The NQF measure evaluation criteria can be viewed as a hierarchy that guides the sequential process for evaluating measures. As described in some of the foundational work for NQF processes:

> If a measure is not important, its other characteristics are less meaningful. If a measure is not scientifically acceptable, its results may be at risk for improper interpretation. If a measure is not interpretable [usable] we probably do not care if it is feasible. If a measure is not feasible, alternative approaches to acquiring important information should be considered. (p. I-40)[1]

Once a measure has been determined to meet the criterion of *Importance to Measure and Report*, it is evaluated on the criterion of *Scientific Acceptability of Measure Properties*. This criterion addresses the basic measurement principles of reliability and validity. The NQF evaluation criteria parallel best practices for measure development, which include testing reliability and validity.[2, 3]

NQF's measure evaluation criteria include a variety of types of evidence as indicated in Table 1. The criterion, *Scientific Acceptability of Measure Properties*, addresses *how* the healthcare quality concept is measured. This criterion includes reliability (2b) and validity (2c), as well as precision of specifications (2a) and potential threats to valid conclusions about quality related to exclusions (2d), risk adjustment for outcome and resource use measures (2e), and comparability of results from different data sources (2g). The other subcriteria include identification of differences in performance (2f) and specifications to detect disparities (2h).

Table 1: Measure Evaluation Criteria and Type of Evidence

Evaluation Criteria	Type of Evidence
1. Importance to measure and report 1a. High impact 1b. Opportunity for improvement 1c. Evidence that supports the focus of measurement	Epidemiologic data Resource use data Health services research Clinical research
2. Scientific acceptability of measure properties 2a.-2g. Reliability, validity, risk adjustment	Psychometric testing—reliability and validity, adequacy of risk-adjustment, etc.
3. Usability 3a. Demonstration of understanding and usefulness for public reporting and quality improvement	Data and/or qualitative information demonstrating usefulness for public reporting and quality improvement
4. Feasibility 4e. Demonstration the measure can be implemented	Data and/or qualitative information demonstrating the measure can be implemented

Reliability and Validity

A quality measure is a numeric quantification of the relatively abstract construct of quality of healthcare, which is measured imperfectly. The concepts of reliability and validity can be applied to the individual data elements used in a measure (e.g., diagnosis, medication, admission date, birth date), as well as the computed performance measure score (e.g., rate, proportion, average).

Reliability refers to the *repeatability or precision of measurement*. Reliability of data elements refers to repeatability and reproducibility of the data elements for the same population in the same time period. Reliability of the measure score refers to the proportion of variation in the performance scores due to systematic differences across the measured entities (or signal) in relation to random error (or noise).

Validity refers to the *correctness of measurement*. Validity of data elements refers to the correctness of the data elements as compared to an authoritative source. Validity of the measure score refers to the correctness of conclusions about the quality of measured entities that can be made based on the measure scores (i.e., a higher score on a quality measure reflects higher quality).

A measure score is an approximation of a theoretical "true" score plus error: The more error, the less reliable and valid is the measurement. Random or chance errors affect the reliability or repeatability of measurement, and systematic errors affect the validity or correctness of the conclusions one can make based on the measure score. Threats to reliability include ambiguous measure specifications (including definitions, codes, data collection, and scoring) and small case volume or sample size. Threats to validity include other aspects of the measure specifications such as inappropriate exclusions, lack of appropriate risk adjustment or risk stratification for outcome and resource use measures, use of multiple data sources or methods that result in different scores and conclusions about quality, and systematic missing or "incorrect" data. Most importantly, a measure may be invalid because the measurement has not correctly captured the concept of quality that it was intended to measure.

Reliability and validity are not all-or-none properties; rather, measures of reliability and validity produce graduated results that always require interpretation. Furthermore, reliability and validity are not static; they are influenced by the conditions under which the measures are implemented (e.g., local documentation and coding practices, structures of records, etc.). Evidence of validity, in particular, is accumulated over time. A discussion of measurement concepts can be accessed in an online research methods knowledge base.[4] Rubin et al.[3] and others[5] provide examples of reliability and validity testing in quality measure development. Over the past four to five decades numerous methods have been devised to test measures and thus address the measure properties inherent to all measurement. These approaches provide empirical evidence of the properties of reliability and validity. Examples of approaches to reliability and validity testing can be found in Tables A-1 through A-5 in Appendix A and in the literature.[6, 7]

Reliability is often considered to be necessary, but not sufficient, for achieving validity. That is, if a measure is not reliable, then a valid conclusion about quality will not be possible. Furthermore, a measure may be reliable but lead to incorrect (invalid) conclusions. However, this relationship between reliability and validity is not universally held[8, 9] and may depend on how a measure is defined. For example, for a patient-level measure of systolic blood pressure (BP), the mean of multiple readings could be accurate even though the individual BP readings are unreliable (i.e., with substantial random error). However, for a patient-level measure defined as one systolic BP over 140, then error or unreliability of the BP reading could lead to categorizing the BP incorrectly as either over or under 140 and hence loss of validity.

Evaluation of the scientific acceptability of a measure does not occur in a vacuum. The Task Force was aware of factors within the current environment related to performance measurement and that the recommendations would have implications for both measure developers and healthcare providers. As the stakes around quality measurement for accountability and payment increase, the potential for conflicts among various stakeholder perspectives also increases. For example, some observers have suggested that existing measure evaluation criteria are too stringent (allowing “the perfect to be the enemy of the good”); while others have suggested that the criteria are not stringent enough. Some contend that healthcare providers use adherence to the criteria for scientific acceptability of measure properties as a barrier to making performance information available; others maintain that unless a measure has adequate measurement properties it cannot provide useful information. Nonetheless, the consequences of using unreliable or invalid measures can at times be significant for those being measured as well as for those using measures to select healthcare providers. Resources may be wasted or misdirected, and unreliable and invalid measures may result in patients being misinformed, misdirected, or subjected to unintended harmful consequences. The Task Force therefore made a deliberate attempt to put forth recommendations that balance the goal of endorsing measures that are sufficiently reliable and valid to make them meaningful and able to minimize unintended consequences with testing requirements that are not so high as to stifle measure development and innovation.

Reporting of Measure Scores and Scientific Acceptability

NQF does not determine the specific use or reporting formats of the measures it endorses. Nonetheless, the confidence in a measure can be related to the context in which the measure is used and the choices made in reporting performance measure scores. For example, researchers at RAND[10]demonstrated that the number of categories chosen for performance reporting (e.g., high/low or high/medium/low) influences the likelihood of misclassification, which, by definition, is invalid reporting of performance. Reporting performance from highest to lowest, without information about margin of error and meaningful differences, limits and may misrepresent the knowledge to be gained from measures. However, confidence intervals or other technical explanations may render the information incomprehensible to some audiences. Combining measures into a composite may simplify reporting, make the metrics more usable for consumers, and provide another way for providers to view performance. Yet, composite measures also have the potential to be misleading depending on the component measure and methods used to combine results.[11] Finding the right balance is important. Because NQF endorsement does not dictate how measures are used, the Task Force was not asked to make recommendations on reporting, but these issues are highlighted for further consideration.

Measure Testing Issues Identified with Measures Submitted to NQF

The Task Force understood its charge as emerging from several years of NQF experience with measure evaluation. This experience, enumerated below in six points, informed the Task Force's recommendations. First, the NQF portfolio of endorsed measures shows considerable variation in the level of rigor used in measure testing. Measure developers are expected to address testing requirements in a way that is most appropriate and feasible for the measure and data source involved. Nonetheless, some developers submit limited information on reliability or validity testing, perhaps because of a lack of expertise or resources. In contrast, other developers conduct formal reliability and validity testing and demonstrate that a proposed measure generates reproducible results and credible conclusions about quality.

Second, when reliability and validity testing results have been submitted, there has been variability in the scope of testing and the rigor of methods and statistical analysis. For example, developers may assess reliability of categorical data elements using the percentage of agreement between raters or the kappa statistic, which adjusts for chance agreement. In some cases, testing may be conducted with a particular data source, such as the paper medical record, even though the measure is specified using a different data source, such as the electronic health record.

Third, there has been some confusion regarding what is considered testing of scientific acceptability. Terms such as “measure testing,” “pilot testing,” and “field testing” are commonly used in the discipline of measure development and include reliability and validity testing, as well as other aspects of measure development such as feasibility analysis. For example, measure submissions may include descriptive statistics that demonstrate that the data are available and can be analyzed to produce scores, but do not specifically address reliability or validity.

Fourth, some submissions rely on an assumption of reliability and validity rather than providing empirical evidence. This assumption may be based on prior use of the measure or some aspects of the measure specifications (e.g., diagnosis codes are relatively well defined and used in accordance with coding rules). In some cases an argument is made that a data source would become more reliable and valid if a quality measure was implemented and publicly reported.

Fifth, measure developers rarely submit analyses justifying exclusions or demonstrating comparability of different methods of data collection.

Sixth, steering committees may variably weigh the strengths and weaknesses of the evidence for reliability and validity in their recommendation for endorsement. In summary, although NQF has been raising the bar of expectations and has been introducing greater rigor and standardization to the evaluation process, the NQF portfolio of endorsed measures still includes varying levels of methodological rigor.

Electronic Health Records and Electronic Measures

Development and implementation of electronic health record (EHR) systems hold great promise for the efficient collection of clinical data that can be used for quality measurement. National initiatives call for the adoption of EHRs that include the capability for quality measurement, and NQF has made endorsing quality measures specified for EHRs an important goal. Data stored in

EHRs facilitate reporting of quality measures because EHR data 1) are clinically specific, 2) include a large variety of data types including physiologic data such as laboratory values, and 3) decrease the burden of data collection through automated identification, extraction, computation, and aggregation.

Although the concepts of reliability and validity apply equally to measures derived from EHRs, the EHR presents additional issues related to measure testing. Widespread EHR data are not yet available for measure development and testing. In addition, because there are numerous EHR vendors and home-grown EHR systems, it can be difficult to insure that the selected data fields of interest for any particular measure are comparable among different EHRs. Recommendations regarding testing and evaluation of EHR measures are addressed in Section III.

Summary of Background

- There are no perfect quality performance measures, and there will be some error in all measurement. Performance measurement science is an imperfect science.
- Measurement principles of reliability and validity apply to quality performance measures regardless of data source.
- Reliability and validity are not all-or-none properties and involve a matter of degree.
- Reliability and validity are not static properties and can vary under the conditions of implementation.
- Reliability and validity can apply to individual data elements used in a measure, as well as the computed measure score.
- Reliability does not guarantee validity.
- Variability in measure scores that is attributable to either random error (noise) or systematic error (biased measurement) is misleading and leads to unwarranted conclusions about quality.
- NQF is ultimately concerned with endorsing measures that produce scores from which valid (i.e., correct) conclusions about the quality of care can be made.

- A measure that is not a valid indicator of quality is not useful for making decisions about selecting healthcare providers based on quality or for investing time and resources into improvement.

RECOMMENDATIONS

The recommendations in this report are intended to provide additional guidance and clarification regarding the NQF criteria related to measure testing and scientific acceptability. However, the guidance does not address the unique aspects of testing for composite measures as indicated in the composite measure evaluation criteria. The guidance is not intended to provide a detailed primer on methods for measure testing or a definitive scoring system for measure evaluation. Evaluation still requires judgment regarding the adequacy of the empirical testing evidence. The recommendations should promote greater consistency in applying the NQF criteria while maintaining consideration of multi-stakeholder perspectives during the evaluation. This guidance then replaces any previous guidance on measure testing (e.g., “field” testing requirements in the time-limited endorsement policy).

I. Recommendations for Empirical Evidence of Reliability and Validity

Before developing guidance on the specific testing criteria, the Task Force was asked to consider a fundamental question of whether reliability and validity need to be demonstrated empirically or can be assumed or agreed upon through various review or consensus processes. The Task Force recommended that ***empirical evidence of reliability and validity should be expected for all measures endorsed by NQF***.

Rationale for Empirical Evidence

Although reliability and validity are not static properties and can vary under different conditions of implementation (e.g., local documentation and coding practices, structures of paper or electronic records, etc.), the purpose of reliability and validity testing for NQF endorsement is to demonstrate that a measure can be reliable and valid when implemented as specified. Although precise specifications provide a foundation for consistent implementation and thus increase the likelihood of reliability, reliability cannot be assumed. Evidence for the measure focus (NQF criterion 1c) provides a foundation for the validity of the measure as an indicator of quality, but the way a measure is specified can affect the validity of the conclusions about quality.

Implementation and reporting of measures is expected to lead to improvements in documentation, data coding, and data capture and thus reliability and validity. This assumption of improved reliability and validity over time applies to all measures regardless of data type; however, it does not negate the need for empirical demonstration of reliability and validity when a measure is being considered for endorsement.

Recommendations for measures specified for EHRs are addressed in a separate section (Section III), because they are newer than measures based on other data types and the EHR data has some unique features. For example, the clinician is often the source of data in the EHR, and the data are intended for use in care management. However, these distinctions are not absolute, and the requirement for demonstrating scientific acceptability of measure properties applies equally to EHR measures and measures based on other data types. Administrative claims data and EHR data may be viewed as complementary sources of information, each with their own strengths and limitations.

Strategies to Mitigate the Burden of Testing

Although the Task Force was clear about requirements for empirical evidence of reliability and validity, it also recognized the practical implications of these requirements for measure developers. Therefore, the Task Force further recommended some strategies that could minimize the burden of testing as follows:

- Evidence for reliability and validity may be accumulated over time, and evaluators should remain flexible with regard to the extent of testing evidence submitted. The scope of testing may be on a relatively small scale for initial endorsement, followed by further analyses to support continued endorsement at the time of endorsement maintenance review.
- Reliability and validity testing may be conducted on a sample of the measured entities. The analytic unit of the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.

  - The sample should represent the variety of entities whose performance will be measured. The Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
  - The sample should include adequate numbers of units of measurement *and* adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.
  - When possible, units of measurement and patients within units should be randomly selected.
- Reliability and validity testing may be conducted for *either* the data elements used to calculate the measure score *or* the computed measure score, to achieve an acceptable rating for endorsement. Ideally, testing is conducted for both the critical data elements and the computed measure score, but only one level of testing would be required for endorsement. See Tables A-1 to A-5 in Appendix A for examples of reliability and validity testing of data elements and measure scores.
- Separate reliability testing of the *data elements* is not required if empirical validity testing of the data elements (see Table A-4) is conducted (e.g., if the validity of ICD-9 codes in administrative claims data as compared to clinical diagnoses in the medical record is demonstrated, then inter-coder or inter-abstractor reliability would not be required).
- Prior evidence of reliability or validity of data elements (see Tables A-2 and A-4 in Appendix A) for the data type specified in the measure (e.g., hospital claims) can be used as evidence for those data elements. Prior evidence could include published or unpublished testing that:
  - includes the same data elements; and
  - uses the same data type (e.g., claims, chart abstraction, etc.); and
  - is conducted on a sample as described above (i.e., representative, adequate numbers, and randomly selected, if possible).
- Because validity testing of measure scores can be quite burdensome, a formal and systematic testing of face validity as described in Table A-3 could be acceptable for a moderate rating of measure score validity.[12, 13] Key components of acceptable face

validity include a systematic and transparent process, the inclusion of identified experts, and explicit discussion of whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The process and results need to be presented.

The Task Force further acknowledged that there are degrees of reliability and validity and the following guidance distinguishes ideal testing and evidence from what is acceptable for endorsement by NQF. Measures without empirical testing of reliability and validity should be considered untested measures and subject to NQF's conditions for considering untested measures for endorsement. Untested measures are addressed in Section IV.

II. Recommendations for the Type of Testing and Results Needed to Demonstrate Scientific Acceptability of Measure Properties

How should participants in the evaluation process assess the evidence provided when measures are submitted? The Task Force chose to provide guidance on measure testing through the development of rating categories for the reliability and validity of measures being considered for endorsement. This approach requires well-defined descriptions of the rating scheme to reduce ambiguity and miscommunication. Although the Task Force has tried to achieve this precision, it recognizes that inevitably there will be some ambiguity and room for interpretation. In addition, the rating descriptions provided in this report may require further clarification and/or revision. Finally, the Task Force was not able to fully assess the impact of the proposed rating system on the measure endorsement process. Therefore, this proposed approach to evaluating scientific acceptability of measure properties should be monitored to ensure that it achieves the intent of endorsing reliable and valid measures and does not unduly impede endorsement of measures.

The Task Force chose to provide guidance on evaluating *Scientific Acceptability of Measure Properties* using a two-step process. First, guidance is provided on how to rate the evidence for reliability and validity. Second, guidance is provided on how to use the ratings to determine if the criterion of *Scientific Acceptability of Measure Properties* is met.

Table 2 provides the guidance for rating the level of evidence for reliability and validity, which is classified as high, moderate, low, or inadequate. The ratings depend on the level of testing conducted, appropriateness of the selected method, scope of testing, and testing results meeting acceptable norms. This table applies to all types of measures and data types; however, the rating scale in Table 4 applies specifically to measures specified for EHRs.

The rating scheme is structured around a distinction between testing the data elements used to calculate a measure (e.g., diagnosis, procedure, age) and testing the computed measure scores (e.g., rate, proportion, average). The data elements are often patient-level information on individual patients (e.g., blood pressure, lab value, medication, surgical procedure, death); the computed measure score represents an aggregation of all the appropriate patient-level data (e.g., proportion of patients who died, average lab value attained) for the entity being measured (e.g., hospital, nursing home, clinician). Some measures rely on many data elements and testing at the data element level does not necessarily need to be conducted for every single data element. Testing should include the critical data elements that contribute most to the computed measure score.

The Task Force determined that it was not possible to set specific statistical thresholds to indicate the degree of reliability and validity that would apply to all situations. Therefore, ***both the moderate and high ratings require results within acceptable norms.*** The distinction between the high and moderate ratings is whether testing is conducted at either the data element level or the computed measure score level (moderate rating) *or* at both levels (high rating). ***The moderate rating is sufficient for passing the criterion and potential endorsement.*** The requirements for the moderate rating provide measure developers with flexibility and minimize the burden of measure development. Table A-6 provides some examples for interpreting statistical results.

Results that are not within acceptable norms indicate unreliable or invalid measurement and would receive the low rating. If the testing was conducted with an inappropriate method or inadequate scope (i.e., representativeness, sample size), then there would be insufficient evidence to evaluate reliability and/or validity and the measure would be considered untested. As noted previously, untested measures would not be rated on reliability and validity, and special considerations for untested measures are addressed in a separate section (see Section IV).

The rating scale presented in Table 2 is not intended to be a definitive scoring system. The determination of adequate testing and results still requires judgment that incorporates a variety of considerations including:

- whether the test was appropriate for the specified measure and purpose of testing;
- whether the scope of testing (i.e., representativeness, sample size) was adequate; and
- whether the results indicate acceptable level of reliability or validity.

Table 2: Evaluation Ratings for Reliability and Validity

Rating	Reliability	Validity
High	All measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring, etc.) are unambiguous and likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.; **AND** Empirical evidence of reliability of **BOTH** data elements (Table A-2) **AND** measure score (Table A-1) within acceptable norms: • Data element: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity (Table A-4) was demonstrated*; **AND** • Measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of **BOTH** data elements (Table A-4) **AND** measure score (Table A-3) within acceptable norms: • Data element: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **AND** • Measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased
Moderate	All measure specifications are unambiguous as noted above **AND** Empirical evidence of reliability within acceptable norms for either critical data elements **OR** measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either critical data elements **OR** measure score as noted above; **OR** Systematic assessment of face validity of measure

		score as a quality indicator (as described in Table A-3) explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased
Low	One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are ambiguous with potential for confusion in identifying who is included and excluded from the target population, or the event, condition, or outcome being measured; or how to compute the score, etc.; **OR** Empirical evidence (using appropriate method and scope) of unreliability for either data elements **OR** measure score, i.e., statistical results outside of acceptable norms	The measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and scope) of invalidity for either data elements **OR** measure score, i.e., statistical results outside of acceptable norms **OR** Identified threats to validity noted above are empirically assessed and determined to bias results
Insufficient Evidence	Inappropriate method or scope of reliability testing	Inappropriate method or scope of validity testing (including inadequate assessment of face validity as noted above); **OR** Threats to validity as noted above are likely and are NOT empirically assessed

Table 3 presents the Task Force's recommendation on how the ratings for reliability and validity are used to determine whether a measure adequately meets the criterion of *Scientific Acceptability of Measure Properties*. Moderate ratings for both validity and reliability as described in Table 2 (and Table 4) would be required to pass this criterion and to be acceptable for endorsement. A high rating is not required for endorsement, but it represents current thinking about best practices in measure development. **A measure that does not pass the criterion of *Scientific Acceptability of Measure Properties* would not be recommended for endorsement.**

Table 3: Evaluation of Scientific Acceptability of Measure Properties Based on Reliability and Validity Ratings

Validity Rating	Reliability Rating	Pass *Scientific Acceptability of Measure Properties* for Initial Endorsement*	
High	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Low	Any rating	No	Validity of conclusions about quality is the primary concern. If evidence of validity is rated low, the reliability rating will usually also be low. Low validity and moderate-high reliability represents inconsistent evidence.

*A measure that does not pass the criterion of *Scientific Acceptability of Measure Properties* would not be recommended for endorsement.

Tables A-1 through A-5 in Appendix A present some common approaches to testing reliability and validity for data elements and the computed measure score that can be applied to quality performance measures. Measure developers should select the testing that is appropriate and feasible for the measure under consideration and that will at least meet the moderate rating as described in Table 2. Table A-5 addresses potential testing and analysis related to the threats to validity represented by other subcriteria under *Scientific Acceptability of Measure Properties*. Measure developers should identify the potential threats to validity for the specific measure and should conduct analyses to demonstrate that the results are not biased. Information on interpretation of the common statistical analyses used to demonstrate reliability and validity is provided in Table A-6; however, those norms provide only general guidelines, and testing results must be interpreted within the unique context of the specific measure.

The information on approaches to testing in the Appendix is not meant to be prescriptive or exhaustive. Other approaches may be used if they employ appropriate rationale and methods. For example, if consistency of data or measure scores between two time periods or test/retest is proposed as a test of reliability, then the rationale for expecting stability (rather than change) over the time period is important to discuss. Calculation of measure scores and descriptive statistics, or the fact that a measure has been in use, do not constitute empirical evidence of reliability or validity. Such information may be relevant to the subcriteria of opportunity for improvement (1b), identification of differences in performance (2f), usability of the measure

(3a), and feasibility of implementation (4e), but descriptive data alone does not address the reliability or validity of the measure.

III. Recommendations for Measures Specified for EHRs

The EHR holds significant promise for improving the measurement of healthcare quality. The availability of a broad range of reliable and valid data elements for quality measurement without the burden of data collection is widely anticipated. Because clinical data can be entered directly into standardized computer readable fields, the EHR will be considered the authoritative source of clinical information. Quality measures based on EHRs use clinical information recorded by healthcare clinicians in discrete computer readable fields; therefore, measurement errors due to manual abstraction, coding by persons other than the originator, or transcription could be eliminated. Despite these potential advantages over current data sources, several potential sources of error pose threats to the reliability and validity of data elements and computed measure scores for EHR measures including: 1) incorrect measure specifications, including code lists, logic, or computer readable programming language; 2) EHR system structure or programming that does not comply with standards for data fields, coding, or exporting data; 3) difference in use of data fields by different users or entry into the wrong EHR field; 4) entry of incorrect information; and 5) incorrect parsing of data by natural language processing software used to analyze information from text fields. All of these potential errors are analogous to sources of error with measures based on other data sources.

Table 4 provides the guidance for rating the level of evidence for reliability and validity of EHR measures, and it is analogous to the ratings in Table 2. Table 3 indicates how the ratings are used to make a determination if the *Scientific Acceptability of Measure Properties* criterion has been met for EHR measures. Approaches to testing the reliability and validity of the EHR measure score are the same as for any measure as noted in Tables A-1 and A-3.

Tables 2 and 4 differ in two ways. First, EHR measures must be specified in accordance with the Quality Data Model (QDM, formerly called the QDS).[14] The reason for requiring specifications using the QDM is twofold: 1) the QDM can be translated to computer-readable specifications that can be applied to EHRs; and 2) the structure of the QDM will help fulfill the criterion for precise specifications. The QDM will be updated on a regular basis; therefore, if a measure needs a quality data element that is not currently available, then there will be a process to consider additional quality data elements so that the measure could achieve a moderate or high rating.

Second, data elements for quality measures, which are extracted from EHRs using computer programming, are by virtue of automation repeatable (reliable); however, they can be wrong (invalid). Different uses of an EHR data field by clinicians or different data processing or extraction protocols in different EHRs can result in incorrect or missing data and produce different performance scores. Therefore, testing at the data element level should focus on validity as discussed below. Focusing on validity testing of data elements is consistent with the rating system for all measures presented in Table 2—that is, if empirical validity testing of the data elements is conducted, then separate reliability testing of the data elements is not required.

An approach to testing the validity of data elements analyzes the agreement between data elements and scores obtained with data exported electronically using the EHR measure specifications to those obtained by review and abstraction of the *entire* EHR, preferably using EHRs that comply with standards. This approach has been reported in the literature[15-17] and by HealthPartners in a Commonwealth Fund report[18] on performance measures and EHRs. As with measures for other data types, testing may be conducted on a sample of the measured entities (see Section I).

Because EHR databases may not be available for such testing, another approach is to apply the EHR measure to a simulated data set that reflects standards for EHRs and includes sample patient data with the elements needed for the specified measure. Because the simulated data set is constructed, the values for the data elements and scores are known. When the EHR specifications are applied to the simulated data set, they should return the known values of the data elements and scores.

With either approach, when the results obtained for the EHR measure do not match the known values in the simulated data set or the abstracted data, an analysis is conducted to determine the source of error. If the error is related to the measure specifications, including code lists, logic, and computer readable programming language, then it would be corrected before submission for endorsement. If the source of error is due to clinical data entry practices and EHR structures unique to specific organizations, then the error would not be mitigated by changes to the EHR measure specifications, but it could indicate the need for further evaluation of feasibility and for alternative data fields.

The recommended approach to evaluating reliability and validity of data elements for EHR measures accounts for the current environment in which standards for EHRs and EHR measures are under development and have not yet been widely adopted. Therefore, testing sites are limited, and testing in a sample of EHR systems may not be representative of all systems. However, this is no different from testing the data elements for measures based on other data sources in a sample of the measured entities whose data practices may vary. As noted in the Background, reliability and validity are not static properties, and no one test is definitive.

Measure testing requirements should not impede the adoption of EHRs and EHR measures, but they should be true to the principles of scientific acceptability of measure properties. EHRs and EHR measures are new and will most likely require some adjustment of local EHR structures and recording practices to meet standards. Therefore, providers should be encouraged to conduct their own internal reliability studies.

Previously endorsed measures specified for chart abstraction or administrative claims data may be appropriate for re-specification for EHRs. Although these endorsed measures should have already been tested for reliability and validity, the EHR measure specifications must be assessed for similarity to the original specifications, which also is addressed in Table 4. In some cases, the EHR specifications will represent a substantive change to the measure so that an assessment of reliability and validity of the EHR measure also is needed.

Table 4: Evaluation of Reliability and Validity of Measures Specified for EHRs

	New Measure Specified for EHR		**Modifications for Endorsed Measure *Re-specified* for EHRs**
Rating	**Reliability Description and Evidence**	**Validity Description and Evidence**	
High	All EHR measure specifications are unambiguous[+] and include only data elements from the Quality Data Model (QDM)* including quality data elements, code lists, and measure logic; **OR** new data elements are submitted for inclusion in the QDM; **AND** Empirical evidence of reliability of both data element **AND** measure score within acceptable norms: • Data element: reliability (repeatability) assured with computer programming—**must test data element validity** **AND** • Measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) reflect the quality of care problem (1a,1b) and evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of both data elements **AND** measure score within acceptable norms: • Data element: validity demonstrated by analysis of agreement between data elements electronically extracted and data elements visually abstracted from the entire EHR with statistical results within acceptable norms; **OR** complete agreement between data elements and computed measure scores obtained by applying the EHR measure specifications to a simulated test EHR data set with known values for the critical data elements; **AND** • Measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased	The EHR measure specifications use only data elements from the Quality Data Model (QDM)* and include quality data elements, code lists, and measure logic; **AND** Crosswalk of the EHR measure specifications (QDM quality data elements, code lists, and measure logic) to the endorsed measure specifications demonstrates that they represent the original measure, which was judged to be a valid indicator of quality; **AND** Analysis of comparability of scores produced by the retooled EHR measure specifications with scores produced by the original measure specifications demonstrated similarity within tolerable error limits
Moder-ate	All EHR measure specifications are unambiguous[+] and include only data elements from the QDM;* **OR** new data elements are submitted for inclusion in the QDM; **AND** Empirical evidence of reliability within acceptable norms for either data elements **OR** measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either data elements **OR** measure score as noted above; **OR** Systematic assessment of face validity of measure score as a quality indicator (as described in Table A-3) explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased	The EHR measure specifications use only data elements from the QDM as noted above **AND** Crosswalk of the EHR measure specifications as noted above demonstrates that they represent the original measure **AND** For measures with time-limited status, testing of the original measure and evidence ratings of moderate for reliability and validity as described in Table 2.
Low	One or more EHR measure specifications are ambiguous[+] or do not use data elements from the QDM*; **OR** Empirical evidence of unreliability for either data elements **OR** measure score—i.e., statistical results outside of acceptable norms	The EHR measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and scope) of invalidity for either data elements **OR** measure score— i.e., statistical results outside of acceptable norms **OR** Identified threats to validity noted above are empirically assessed and determined to bias	The EHR measure specifications do not use only data elements from the QDM; **OR** Crosswalk of the EHR measure specifications as noted above identifies that they do not represent the original measure **OR** For measures with time-limited status, empirical evidence of low

	New Measure Specified for EHR		Modifications for Endorsed Measure *Re-specified* for EHRs
Rating	**Reliability Description and Evidence**	**Validity Description and Evidence**	
		results	reliability or validity for original time-limited measure
Insufficient evidence	Inappropriate method or scope of reliability testing	Inappropriate method or scope of validity testing (including inadequate assessment of face validity as noted above) **OR** Threats to validity as noted above are likely and are NOT empirically assessed	Crosswalk of the EHR measure specifications as noted above was not completed OR For measures with time-limited status, inappropriate method or scope of reliability or validity testing for original time-limited measure

[+] Specifications are considered unambiguous if they are likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.

*QDM (formerly called the QDS) elements should be used when available. When quality data elements are needed but are not yet available in the QDM, they will be considered for addition to the QDM.

IV. Recommendations Related to Untested Measures

Measures without empirical evidence of reliability and validity are considered untested. Untested measures may be eligible for time-limited endorsement if all of the following conditions are met:

- the measure's specific topic of interest has not been addressed by an endorsed measure;
- a critical timeline must be met (e.g., legislative mandate);
- the measure is not complex (e.g., composite, requires risk adjustment); and
- the developer can complete testing within 12 months.

In addition to passing the *Importance to Measure and Report* criterion, untested measures must demonstrate an adequate foundation for both reliability and validity as described in Table 5. That is, measures should be fully and precisely specified and be consistent with the evidence provided. Measures that do not meet these minimum requirements are not ready for testing and should not be recommended for time-limited endorsement.

Table 5: Minimum Requirements for Untested Measures under *Scientific Acceptability of Measure Properties*

Foundation for Reliability	Foundation for Validity
All measure specifications (e.g., numerator, denominator, exclusions, scoring) are unambiguous and likely to consistently 1) identify who is included and excluded from the target population; 2) identify the process, condition, event, or outcome being measured; 3) compute the measure score; etc. All EHR measure specifications are unambiguous and include only data elements from the quality data set (QDM)***** including quality data elements, code lists, and measure logic, **OR** new data elements are submitted for inclusion to the QDM.	The measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) reflect the quality of care problem (1a, 1b) and evidence cited in support of the measure focus (1c) under *Importance to Measure and Report.*

*QDM (formerly called the QDS) elements should be used when available. When quality data elements are needed but are not yet available in the QDM, they will be considered for addition to the QDM.

V. Recommendations for Testing Required for Maintenance of Endorsement

The above guidance on testing and evidence of reliability and validity for initial endorsement also applies to endorsement maintenance, with a few modifications. With the NQF system of endorsement cycles, endorsed measures are reviewed for maintenance of endorsement every three years along with new measures. Both new and endorsed measures will be required to meet the measure evaluation criteria, including reliability and validity.

The Task Force agreed that reliability and validity should be evaluated when measures are reviewed for maintenance of endorsement. Several considerations were relevant to the deliberations on this subject, including: recognizing that reliability and validity are not static properties, no one test is definitive, evidence accumulates over time, and the proposed rating system permits endorsement of measures that have limited evidence of reliability and validity (moderate rating). However, developers cannot be expected to monitor both reliability and validity indefinitely once these measure properties have been well established.

As outlined in Table 6, at the time of endorsement maintenance review, reliability and validity testing should a) use data from implementation of the endorsed measure as specified and b) focus on the measure score rather than on the data elements. Of particular relevance to a measure in use is information on the accuracy of any classification based on the measure results. If an endorsed measure has not been implemented, then expanded testing in terms of scope and levels is required. The rating system presented in Tables 2 and 3 also applies to the maintenance review. As with initial endorsement, all the other criteria also will be used to determine whether a measure warrants continued endorsement.

Table 6: Scope of Testing Required at the Time of Review for Endorsement Maintenance

	First Endorsement Maintenance Review	Subsequent Reviews
Reliability	**Measure In Use** • Analysis of data from entities whose performance is measured • Reliability of measure scores (e.g., signal to noise analysis) **Measure Not in Use** • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that reliability achieved a high rating
Validity	**Measure in Use** • Analysis of data from entities whose performance is measured • Validity of measure score for making accurate conclusions about quality • Analysis of threats to validity **Measure Not in Use** • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that validity achieved a high rating

VI. Recommendations for Modifications to the NQF Evaluation Criteria

The recommendations of the Task Force as described above resulted in some wording changes to the NQF measure evaluation criteria presented in Table 7, but the intent remains unchanged. Criterion 2, *Scientific Acceptability of Measure Properties*, is primarily about reliability and validity and threats to reliability and validity. This criterion can be simplified by focusing on the concepts of reliability and validity and arranging the subcriteria to reflect their relationship to reliability or validity as follows.

2a. Reliability

2a1. Precise specifications (previously 2a) including exclusions (previously 2d)

2a2. Reliability testing (previously 2b)—data elements or measure score

2b. Validity

2b1. Specifications consistent with evidence (new)

2b2. Validity testing (previously 2c)—data elements or measure score

2b3. Justification of exclusions (previously 2d)—relates to evidence

2b4. Risk adjustment (previously 2e)

2b5. Identification of differences in performance (previously 2f)

2b6. Comparability of data sources/methods (previously 2g)

2c. Disparities (previously 2h)

Table 7: Current and Modified Measure Evaluation Criteria

Current Measure Evaluation Criteria	Modified Measure Evaluation Criteria
2. Scientific acceptability of the measure properties: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.	**2. Scientific acceptability of the measure properties:** Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.
[***See footnotes below the criteria*** footnotes do not begin at 1 because of footnotes related to the first criterion]	[***See footnotes below the criteria*** footnotes do not begin at 1 because of footnotes related to the first criterion]
2a. The measure is well defined and precisely specified[6] so that it can be implemented consistently within and across organizations and allow for comparability. The required data elements are of high quality as defined by NQF's Health Information Technology Expert Panel (HITEP). [7]	**2a. Reliability** **2a1.** The measure is well defined and precisely specified[6] so that it can be implemented consistently within and across organizations and allow for comparability. EHR measure specifications are based on the Quality Data Model (QDM).[7]
2b. Reliability testing[8] demonstrates the measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period.	**2a2.** Reliability testing[8] demonstrates that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or the measure score is precise.
2c. Validity testing[9] demonstrates that the measure reflects the quality of care provided, adequately distinguishing good and poor quality. If face validity is the only validity addressed, it is systematically assessed.	**2b. Validity** **2b1.** The measure specifications[6] are consistent with the evidence presented to support the focus of measurement under criterion 1c. The measure is specified to capture the most inclusive target population indicated by the evidence and exclusions are supported by the evidence.
2e. For outcome measures and other measures (e.g., resource use) when indicated: • an evidence-based risk adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care[11,13] **OR** • rationale/data support no risk adjustment.	**2b2.** Validity testing[9] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. **2b3.** Exclusions are supported by the clinical evidence, otherwise they are supported by evidence[10] of sufficient frequency of occurrence so that results are distorted

Current Measure Evaluation Criteria	Modified Measure Evaluation Criteria
2f. Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful[14] differences in performance **2g.** If multiple data sources/methods are allowed, there is demonstration they produce comparable results. **2d.** Clinically necessary measure exclusions are identified and must be: supported by evidence[10] of sufficient frequency of occurrence so that results are distorted without the exclusion; AND • a clinically appropriate exception (e.g., contraindication) to eligibility for the measure focus [11]; **AND** • precisely defined and specified: – if there is substantial variability in exclusions across providers, the measure is specified so that exclusions are computable and the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); – if patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that it strongly impacts performance on the measure and the measure must be specified so that the information about patient preference and the effect on the measure is transparent[12] (e.g., numerator category computed separately, denominator exclusion category computed separately). **2h.** If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender); OR rationale/data justifies why stratification is not necessary or not feasible. **Footnotes** **6** Measure specifications include the target population (e.g., denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (e.g., numerator), measurement time window, exclusions, risk adjustment, definitions, data elements, data source and instructions, sampling, scoring/computation. **7** The HITEP criteria for high quality data include: a) data captured from an authoritative/accurate source; b) data are coded using recognized data standards; c) method of capturing data electronically fits the workflow of the	without the exclusion; **AND** – Measure specifications for scoring include computing exclusions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); **AND** – If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent[12] (e.g., numerator category computed separately, denominator exclusion category computed separately). **2b4.** For outcome measures and other measures when indicated (e.g., resource use): • an evidence-based risk adjustment strategy (e.g., risk models, risk stratification) is specified; is based on factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care;[11,13] and has demonstrated adequate discrimination and calibration **OR** • rationale/data support no risk adjustment/stratification. **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful[14] differences in performance; OR there is evidence of overall less than optimal performance. **2b6.** If multiple data sources/methods are specified, there is demonstration that they produce comparable results. **2c.** If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender); OR rationale/data justifies why stratification is not necessary or not feasible. **Footnotes** **6** Measure specifications include the target population (denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window,

Current Measure Evaluation Criteria	Modified Measure Evaluation Criteria
authoritative source; d) data are available in EHRs; and e) data are auditable. NQF. *Health Information Technology Expert Panel Report: Recommended Common Data Types and Prioritized Performance Measures for Electronic Healthcare Information Systems*. Washington, DC: NQF; 2008. **8** Reliability testing may address the data items or final measure score. Examples of reliability testing include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. **9** Examples of validity testing include, but are not limited to: determining if measure scores adequately distinguish between providers known to have good or poor quality assessed by another valid method; correlation of measure scores with another valid indicator of quality for the specific topic; ability of measure scores to predict scores on some other related valid measure; content validity for multi-item scales/tests. Face validity is a subjective assessment by experts of whether the measure reflects the quality of care (e.g., whether the proportion of patients with BP < 140/90 is a marker of quality). If face validity is the only validity addressed, it is systematically assessed (e.g., ratings by relevant stakeholders) and the measure is judged to represent quality care for the specific topic and that the measure focus is the most important aspect of quality for the specific topic. **10** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers. **11**Risk factors that influence outcomes should not be specified as exclusions. **12** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions. **13** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care such as race, socioeconomic status, gender (e.g., poorer treatment outcomes of African American men with prostate cancer, inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than adjusting out differences. **14** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74% v. 75%) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is	exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, scoring/computation. **7** EHR measure specifications include data type from the QDM (formerly QDS), code lists, EHR field, measure logic, original source of the data, recorder, and setting. **8** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise). **9** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to, testing hypotheses that the measure scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. **10** Examples of evidence that an exclusion distorts measure results include, but are not limited to, frequency of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers. **11** Risk factors that influence outcomes should not be specified as exclusions. **12** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions. **13** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care such as race, socioeconomic status, gender (e.g., poorer treatment outcomes of African American men with prostate cancer, inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than adjusting out differences. **14** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically

Current Measure Evaluation Criteria	Modified Measure Evaluation Criteria
practically meaningful. Measures with overall poor performance may not demonstrate much variability across providers.	significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74% v. 75%) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less than optimal performance may not demonstrate much variability across providers.

VII. Recommendations for Modifications to the Measure Submission Form

The recommendations of the Task Force resulted in modest changes to the information that is requested on the measure submission form as presented in Table 8. The numbering system will need to be adjusted as appropriate for the reorganization of the subcriteria noted above.

Table 8: Current and Modified Measure Submission

Current Measure Submission Items	Modified Measure Submission Items
Measure Specifications (Measure evaluation criterion 2a) **Items 2a.1–2a.38**	**Measure Specifications (Measure evaluation criterion 2a)** **Items 2a.1–2a.38** The recommendations in this report did not indicate specific changes to the submission items for the measure specifications; however, a few suggestions were made to improve clarity.
2a.12. Risk Adjustment Type No risk adjustment necessary analysis by subgroup case-mix adjustment paired data at patient level risk adjustment devised specifically for this measure/condition risk adjustment method widely or commercially available Other (specify)	**2a.12. Risk Adjustment/Stratification Type** No risk adjustment/stratification is necessary—measure is not an outcome or resource use measure No risk adjustment/stratification is necessary—rationale and analysis provided in Section 2e Stratification/analysis by subgroup—see variables in 2a.11 Statistical risk model—specifications 2a.14 Other (specify)
2a.14. Risk Adjustment Methodology/Variables *(List risk adjustment variables and describe conceptual models, statistical models, or other aspects of model or method)*	**2a.14. Specifications for Statistical Risk Model and Variables Included** *(Name the statistical method (e.g., logistic regression) and list the risk model variables, all definitions, and codes with descriptors. Development and testing are reported in Section 2e.)*
2a.21. Calculation Algorithm *(Describe the calculation of the measure as a flowchart or series of steps)*	**2a.21. Measure Score Calculation Algorithm** *(Describe the calculation of the measure score as a series of steps, including identification of denominator, exclusions, identification of numerator, stratification or adjustment, and classification category.)* **2a.21.1. Measure Algorithm or Flow Diagram** *(Please*

Current Measure Submission Items	Modified Measure Submission Items
	provide a web page URL or attachment. NQF strongly prefers URLs. Attach documents only if they are not available on a web page, and keep attached file to 5 MB or less.)
2a.22. Describe the method for discriminating performance *(E.g., significance testing)*	**2a.22.** Delete as a specification.
Reliability Testing (Measure evaluation criterion 2b) **2b.1. Data Sample** *(Description of data sample and size)* **2b.2. Analytic Methods** *(Type of reliability and rationale, method for testing)* **2b.3. Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*	**Reliability Testing (Measure evaluation criterion 2b)** **2b.1. Data/Sample** *(Description of data/sample and size)* **2b.2. Analytic Methods** *(Method of reliability testing and rationale)* **2b.3. Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*
Validity Testing (Measure evaluation criterion 2c) **2c.1. Data Sample** *(Description of data sample and size)* **2c.2. Analytic Method** *(Type of validity and rationale, method for testing)* **2c.3. Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted)*	**Validity Testing (Measure evaluation criterion 2c)** **2c.1. Data/Sample** *(Description of data/sample and size)* **2c.2. Analytic Method** *(Method of validity testing and rationale; if face validity, describe systematic assessment as)* **2c.3. Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted)*
Measure Exclusions (Measure evaluation criterion 2d) **2d.1. Summary of Evidence Supporting Exclusion(s)** **2d.2. Citations for Evidence** **2d.3. Data Sample** *(Description of data sample and size)* **2d.4. Analytic Method** *(Type of analysis and rationale)* **2d.5. Testing Results** *(E.g., frequency, variability, sensitivity analyses)*	**Measure Exclusions (Measure evaluation criterion 2d)** **2d.1. Summary of Evidence Supporting Exclusion(s)** **2d.2. Citations for Evidence** **2d.3. Data/Sample** *(Description of data/sample and size)* **2d.4. Analytic Method** *(Type of analysis and rationale)* **2d.5. Testing Results** *(e.g., frequency, variability, sensitivity analyses of impact on measure scores)*
Risk Adjustment Strategy (Measure evaluation criterion 2e) **2e.1. Data Sample from Testing or Current Use** *(Description of data sample and size)* **2e.2. Analytic Method** *(Type of risk adjustment, analysis and rationale)* **2e.3. Testing Results** *(Risk model performance metrics)* **2e.4. If outcome or resource use measure is not risk adjusted, provide rationale**	**Risk Adjustment Strategy (Measure evaluation criterion 2e)** **2e.1. Data/Sample** *(Description of data/sample and size used for development and validation)* **2e.2. Analytic Method** *(Description of methods for development and testing of risk model including selection of risk factors)* **2e.3. Testing Results** *(Quantitative assessment of relative contribution of model risk factors; Risk model performance metrics including cross-validation, calibration, and discrimination statistics, and assessment of adequacy in the context of norms for risk models. Provide calibration curve and risk decile plot in attachment.)* **2e.4. If outcome or resource use measure is not risk adjusted, provide rationale for not doing so.**
Identification of Meaningful Differences in Performance (Measure evaluation criterion 2f)	**Identification of Meaningful Differences in Performance (Measure evaluation criterion 2f)**

Current Measure Submission Items	Modified Measure Submission Items
2f.1. Data Sample from Testing or Current Use *(Description of data sample and size)* **2f.2. Methods to Identify Statistically Significant and Practical or Meaningful Differences in Performance** *(Type of analysis and rationale)* **2f.3. Measure Scores from Testing or Current Use** *(Description of scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*	**2f.1. Data/Sample** *(Description of data/sample and size)* **2f.2. Analytic Method to Identify Statistically Significant and Practical or Meaningful Differences in Performance** *(Type of analysis and rationale)* **2f.3. Results** *(Description of measure scores, e.g., distribution by quartile, mean, median, SD, outliers, etc.; identification of statistically significant and meaningfully differences in performance. If no variability, discuss rationale for performance measurement.)*
Comparability of Multiple Data Sources/Methods (Measure evaluation criterion 2g) **2g.1. Data Sample** *(Description of data sample and size)* **2g.2. Analytic Method** *(Type of analysis and rationale)* **2g.3. Testing Results** *(E.g., correlation statistics, comparison of rankings)*	**Comparability of Multiple Data Sources/Methods (Measure evaluation criterion 2g)** **2g.1. Data/Sample** *(Description of data/sample and size)* **2g.2. Analytic Method** *(Type of analysis and rationale)* **2g.3. Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted)*
Disparities in Care (Measure evaluation criterion 2h) **2h.1. If measure is stratified, provide stratified results** *(Scores by stratified categories/cohorts)* **2h.2. If disparities have been reported/identified but measure is not specified to detect disparities, provide follow-up plans**	**Disparities in Care (Measure evaluation criterion 2h)** **2h.1. If measure is stratified to identify disparities, provide stratified results** *(Scores by stratified categories/cohorts)* **2h.2. If disparities have been reported/identified but measure is not specified to detect disparities, provide follow-up plans**

NOTES

1. McGlynn EA, Selecting common measures of quality and system performance, *Med Care*, 2003;41(1 Suppl):I39-I47.
2. McGlynn EA, Asch SM, Developing a clinical performance measure, *Am J Prev Med*, 1998;14(3 Suppl):14-21.
3. Rubin HR, Pronovost P, Diette GB, From a process of care to a measure: the development and testing of a quality indicator, *Int J Qual Health Care*, 2001;13(6):489-496.
4. Trochim WMK, Research methods knowledge base, *Web Center for Social Research Methods*, 2006. Available at: www.socialresearchmethods.net/kb/index.php. Last accessed May 2010.
5. Physician Consortium for Performance Improvement, *Measure Testing Protocol for Physician Consortium for Performance Improvement Performance Measures*, Chicago, IL: American Medical Association, 2007.
6. Bhattacharyya T, Freiberg AA, Mehta P, et al., Measuring the report card: the validity of pay-for-performance metrics in orthopedic surgery, *Health Aff (Millwood)*, 2009;28(2):526-532.
7. Schneider EC, Nadel MR, Zaslavsky AM, et al., Assessment of the scientific soundness of clinical performance measures: a field test of the National Committee for Quality Assurance's colorectal cancer screening measure, *Arch Intern Med*, 2008;168(8):876-882.
8. Moss PA, Can there be validity without reliability?, *Educational Researcher*, 1994;23(2):5-12.

9. Salvucci S, Walter E, Conley V, Fink S, Saba M, *Measurement Error Studies at the National Center for Education Statistics*, Washington, DC: U.S. Department of Education, 1997.
10. Adams JL, Mehrotra A, McGlynn EA, *Estimating Reliability and Misclassification in Physician Profiling*, Santa Monica, CA: RAND Corporation, 2010. Available at www.rand.org/pubs/technical_reports/TR863. Last accessed November 2010.
11. Reeves D, Campbell SM, Adams J, et al., Combining multiple indicators of clinical quality: an evaluation of different analytic approaches, *Med Care*, 2007;45(6):489-496.
12. Fitch K, Bernstein SJ, Aguilar MS, et al., *The RAND/UCLA Appropriateness Method User's Manual*, Santa Monica, CA: RAND Health, 2000. Available at www.rand.org/pubs/monograph_reports/MR1269/. Last accessed November 2010.
13. Spertus JA, Eagle KA, Krumholz HM, et al., American College of Cardiology and American Heart Association methodology for the selection and creation of performance measures for quantifying the quality of cardiovascular care, *Circulation*, 2005;111(13):1703-1712.
14. National Quality Forum, *Health Information Technology Expert Panel II - Health IT Enablement of Quality Measurement*, Washington, DC: NQF, 2009.
15. Baker DW, Persell SD, Thompson JA, et al., Automated review of electronic health records to assess quality of care for outpatients with heart failure, *Ann Intern Med*, 2007;146(4):270-277.
16. Persell SD, Wright JM, Thompson JA, et al., Assessing the validity of national quality measures for coronary artery disease using an electronic health record, *Arch Intern Med,* 2006;166(20):2272-2277.
17. Weiner M, Stump TE, Callahan CM, et al., Pursuing integration of performance measures into electronic medical records: beta-adrenergic receptor antagonist medications, *Qual Saf Health Care*, 2005;14(2):99-106.
18. Briggs JB, Kind EA, Awwad S, et al., *Performance Measures Using Electronic Health Records: Five Case Studies*, New York, NY: The Commonwealth Fund, 2008. Report No.: 1132, Available at www.commonwealthfund.org.

APPENDIX A
COMMON APPROACHES TO MEASURE TESTING

Tables A-1 through A-5 provide examples of the various types of reliability and validity testing that ***could*** be performed. The information in the following tables is not meant to be prescriptive or exhaustive. Other approaches to testing that employ an appropriate method and rationale may be used. Measure developers should select the testing that is appropriate and feasible for the measure being developed and that will meet at least the moderate rating as described in Table 2. Likewise, measure developers should identify the potential threats to validity for the specific measure and conduct analyses to demonstrate adequate control.

The rating scheme and following tables are structured around a distinction between testing the data elements (Tables A-2 and A-4) used to calculate a measure and testing the computed measure scores (Tables A-1 and A-3). The data elements are often patient-level information on individual patients (e.g., blood pressure, lab value, medication, surgical procedure, death); the computed measure score represents an aggregation of all the appropriate patient-level data (e.g., proportion of patients who died, average lab value attained) for the entity being measured (e.g., hospital, nursing home, clinician). Table A-5 includes examples of testing related to threats to validity such as patient factors that may affect an outcome measure. Table A-6 includes examples of interpretation of some statistical results.

Table A-1: Examples of Reliability Testing at the Level of the Computed Performance Measure Score

Reliability Testing—Measure Score	
Data	**Aspect of Reliability/Test**
Reliability testing of the computed measure score does not vary by type of data or type of measure. Requires data for the computed measure scores and the individual patient-level data for the measured entities	**Analysis of proportion of variation due to true differences vs. noise or random variation** Analysis of the relative value of variation in measure scores due to signal (i.e., variation between measured entities) versus noise (i.e., variation within measured entities) using statistical analyses such as Analysis of Variance (ANOVA), Intraclass Correlation Coefficient (ICC), or variance components from a multi-level mixed model[19, 20] Monte Carlo simulation to test Bayesian measures[21] Generalizability analysis based on generalizability theory on the sources of variation[22] **Other:** Other methods may be appropriate and rationale for method chosen should be provided

Table A-2: Examples of Reliability Testing at the Level of the Data Elements

Reliability Testing—Data elements	
Separate reliability testing of the data elements is not required if validity testing is conducted on the data elements.	Empirical validity testing of the data elements (see Table A-4) is conducted and demonstrates the data elements are valid.
Prior evidence of reliability of data elements can be used for evidence of reliability of data elements.	Prior evidence could include published or unpublished testing that: • included the same data elements; and • used the same data type; and • was conducted on a sample as described above (i.e., representative, adequate numbers, and randomly selected, if possible).
Data Type	**Aspect of Reliability/Test**
Retrospective chart abstraction (including registry data abstracted retrospectively from medical records)	**Inter-rater reliability between abstractors** Analysis of agreement using appropriate statistical analyses (e.g., kappa, ICC) with 2nd abstractor on each critical data element and computed measure score
Administrative claims data where codes that are used to represent the primary clinical data (ICD, CPT, CPT-II/G)	**Inter-rater reliability between coders** Analysis of agreement using appropriate statistical analyses (e.g., kappa, ICC) with a 2nd coder on each critical data element and computed measure score
Standardized clinical patient information (MDS, OASIS, registry, potentially some aspects of EHRs) collected by an authoritative source concurrently with care delivery (not abstracted, coded, or transcribed by another person)	**Inter-rater reliability between assessors** Analysis of agreement using appropriate statistical analyses (e.g., kappa, ICC) with 2nd assessor on each critical data element and computed measure score.
EHR clinical record information	Data elements obtained with EHR specifications and data exported electronically from EHRs according to standards are repeatable (reliable) when applied to the same population in the same time period. Testing of data elements should focus on validity.
Survey—single items	**Test-retest reliability** Analysis of agreement between two administrations of the same items (time frame long enough so as not to remember and short enough so as not to have changed)
Instrument/scale	If patient scores from an instrument/scale are used in constructing a performance measure, then generally the reliability of the scale has already been tested and documented and can be used as evidence of data element reliability. **Internal consistency reliability (Cronbach's alpha)** Analysis of the extent to which item responses obtained at the same time correlate highly with each other
Other data type	Rationale should be provided for method chosen to demonstrate reliability

Table A-3: Examples of Validity Testing at the Level of the Computed Performance Measure Score

Validity Testing—Measure Score	
Data	**Aspect of Validity/Test**
Validity testing of the computed measure score does not vary by type of data or type of measure. Requires data for the computed measure scores for the measured entities and other data as necessary for the chosen validity study	**Evidence that supports the intended interpretation of measure scores for the intended purpose—making conclusions about the quality of care** Systematic testing of face validity of the measure score as a quality indicator by identified experts, explicitly addressed the question of whether ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** (using a systematic and transparent process, e.g., modified Delphi, formal consensus process, RAND Appropriateness Method[12], ACC/AHA method)[13] with methods and results reported for review. **Criterion Validity:** Studies to assess the correlation of the computed measure score against some criterion determined to be valid. ***Concurrent*—**Correlation with another measure of the same construct measured at the same time ***Predictive*—**Correlation with another measure of the same construct or an outcome measured at some time in the future **Construct Validity:** Studies to assess how the measure performs based on the theory of the construct. ***Contrasted Groups*—**Study to assess the ability of the measure score to distinguish between groups that it should theoretically be able to distinguish ***Convergent*—**Study to examine the degree to which the measure score is similar to (converges on) other measures of the same construct or measures to which it theoretically should be similar ***Discriminative*—**Study to examine the degree to which the measure score is not similar to (diverges from) other measures to which it theoretically should not be similar **Other:** Other methods may be appropriate, and rationale for method chosen should be provided.

Table A-4: Examples of Validity Testing at the Level of Data Elements

Validity Testing—Data elements	
Prior evidence of validity of data elements can be used for evidence of validity of data elements.	Prior evidence could include published or unpublished testing that: • included the same data elements; and • used the same data type; and • was conducted on a sample as described above (i.e., representative, adequate numbers, and randomly selected, if possible).
Data Type	**Aspect of Validity/Test**
Retrospective chart abstraction (including registry data abstracted retrospectively from medical records)	**Validity of data elements abstracted from medical record as compared to some criterion authoritative source of the same data** Analysis of agreement using appropriate statistical analyses (e.g., sensitivity, specificity, positive predictive value, negative predictive value with some other source of the same information considered to be valid (e.g., original data collection such as survey or observation, vital statistics)
Administrative claims data where codes that are used to represent the primary clinical data (ICD, CPT, CPT-II/G)	**Validity of coded data from claims as compared to some criterion authoritative source of the same data** Analysis of agreement using appropriate statistical analyses (e.g., sensitivity, specificity, positive predictive value, negative predicted value[23, 24]) with manual abstraction from the full medical record as the authoritative source
Standardized clinical patient information (MDS, OASIS, registry, potentially some aspects of EHRs) collected by an authoritative source concurrently with care delivery (not abstracted, coded, or transcribed by another person)	**Validity of data elements from standardized assessment instruments as compared to some criterion authoritative source of the same data** Analysis of agreement using appropriate statistical analyses (e.g., sensitivity, specificity, positive predictive value, negative predictive value) with "expert" assessor (conducted at approximately the same time) **Predictive validity as described in Table A-3** (e.g., patient-level assessment item or score predicts a subsequent patient-level outcome of undisputed importance, such as death or permanent disability)
EHR clinical record information	**Validity of data elements extracted from specified fields in EHRs as compared to some criterion authoritative source of the same data** Analysis of agreement using appropriate statistical analyses (e.g., sensitivity, specificity, positive predictive value, negative predictive value) with data elements abstracted from the entire EHR (not just the fields where the data are expected)[15-17] Demonstration of agreement between data elements and scores obtained by applying the EHR measure specifications to a simulated test EHR data set that reflects standards for EHRs and includes sample patient data with known values for the data elements needed for the specified measure and computed measure score
Survey—single items	**Validity of data elements from survey as compared to some criterion authoritative source of the same data** Analysis of agreement using appropriate statistical analyses (e.g., sensitivity, specificity, positive predictive value, negative predictive value) with some other source of the same information considered to be valid (e.g., medical record, vital statistics)
Instrument/scale	If patient scores from an instrument/scale are used in constructing a performance measure, generally the validity of the scale has already been tested and documented and can be used as evidence of data element validity. **Validity of the content of the items in an instrument or scale**

Validity Testing—Data elements	
	Systematic assessment by subject matter experts that the content of the instrument/scale is representative of the domain being measure **Validity of whether the instrument is consistent with the theoretical construct** Confirmatory factor analysis **Criterion or construct validity as described in Table A-3 of the patient-level score** (e.g., patient-level score predicts a subsequent outcome of undisputed importance, such as death or permanent disability)
Other data type	Rationale should be provided for method chosen to demonstrate validity
All data types	**Other aspects of validity** Other methods and aspects of validity (e.g., as described in Table A-3) may be appropriate for some data elements, and rationale for method chosen should be provided.

Table A-5: Examples of Testing Related to Threats to Validity

Threat to Validity	Testing/Analysis
Threat that differences in measure scores are due to differences in severity of conditions of patients served rather than differences in quality (confounding bias)	For outcome and resource use measures, empirical evidence for the adequacy of adjustment for patient factors (analysis of risk factors, discrimination, and calibration of risk models); **OR** evidence that risk adjustment/stratification is not necessary for fair comparisons (patient outcomes do not vary by patient characteristics)
Threat of bias from differences in data type and/or differences in data collection practices (information bias)	If multiple data sources (e.g., medical record and claims) or methods (e.g., mail survey and interview) are specified, empirical evidence that resulting measure scores are comparable (analysis of agreement between scores based on different data sources)
Threat of bias from missing or “incorrect” data or exclusions (selection/attrition bias)	Sensitivity analysis of the impact of missing or “incorrect” data on resulting measure scores (analysis of patterns of missing data; simulate missing data or “incorrect” data, and analyze impact on measure scores) Analyses of frequency of exclusions, sensitivity analyses with and without the exclusion, and variability of exclusions across providers

Table A-6: Examples of Interpretation of Statistical Results

Test	Interpretation
Kappa[25-27] Measure of agreement between two raters that adjusts for chance agreements for categorical data (nominal, ordinal)	Kappa values range between 0 and 1. 0 and are interpreted as degree of agreement beyond chance. By convention, a kappa > .70 is considered acceptable inter-rater reliability, but this depends on the researcher's purpose[28] 0 No better than chance 0.01-0.20 Slight 0.21-0.40 Fair 0.41-0.60 Moderate 0.61-0.80 Substantial 0.81-1.0 Almost perfect[29]
ICC Alternative measure of agreement when more than two raters or for quantitative data (interval, ratio)	ICC values range between 0 and 1.0. Interpretations are similar for kappa noted above. ICC approaches 1.0 only if there is no variance due to raters.
ANOVA or ICC Used for signal-to-noise analysis for estimated mean (or proportion) **—** analysis of variance between the measured entities (signal) to variance within the measured entities (noise)	F test of equality of means for measured entities; F-1 is an estimate of the ratio of signal to noise, and [1-(1/F)] estimates the fraction of total variance that is due to signal (real variation among measured entities), referred to as interunit reliability (IUR). When F is large, IUR is close to 1 indicating almost all signal and no noise. Zaslavsky[30] demonstrated that value of F should be 10 or greater.
Cronbach's alpha Measure of the average correlation of the items comprising a scale or subscale	A widely-accepted cut-off is **.70 or higher**[31] for a set of items to be considered a scale. Some use .75 or .80, while others are as lenient as .60. That .70 is as low as one may wish to go is reflected in the fact that when alpha is .70, the standard error of measurement will be more than half (0.55) a standard deviation[32]
Pearson Correlation Measure of the degree of association (not agreement) between two quantitative variables	Values range from -1 to +1. The squared correlation represents the proportion of variance shared by the two variables (e.g., correlation of 0.5 represents 25% shared variance). Interpretation depends on statistical significance, size of the correlation, and context (e.g., norms for the concepts; physiologic v. psychosocial concepts). Cohen[33, 34] gives the following guidelines for the correlation effect size in the social sciences: 0.10-0.23**—**small 0.24-0.36**—**medium 0.37 or larger**—**large
Spearman (rank order) correlation Measure of the degree of association (not agreement) for rank-order variables	Values range from -1 to +1 A high positive value indicates a strong tendency for the paired ranks to be similar; a negative indicates the paired ranks to be opposite.

APPENDIX B
TASK FORCE MEMBERS

Timothy G. Ferris, MD, Mphil, MPH
(Chair and CSAC Member)
Associate Professor of Medicine and Pediatrics
Massachusetts General Hospital/Institute for Health Policy
Boston, MA

Andy Amster, MSPH
Director, Integrated Analytics
Kaiser Permanente
Los Angeles, CA

Nancy Dunton, PhD
Research Professor
University of Kansas School of Nursing
Kansas City, KS

Steven Findlay, MPH
Senior Health Policy Analyst
Consumers Union
Washington, DC

David S.P. Hopkins, MS, PhD
(CSAC Member)
Director of Quality Measurement
Pacific Business Group on Health
San Francisco, CA

Karen Kmetik, PhD
Vice President for Performance Improvement
American Medical Association convened Physician Consortium for Performance Improvement
Chicago, IL

Rebecca S. Lipner, PhD
Vice President of Psychometrics and Research Analysis
American Board of Internal Medicine
Philadelphia, PA

Jerod Loeb, PhD
Executive Vice President for Research
The Joint Commission
Oakbrook Terrace, IL

Sean O'Brien, PhD
Assistant Professor, Dept. of Biostatistics and Bioinformatics
Duke University Medical Center
Durham, NC

Patrick S. Romano, MD, MPH
Professor of Medicine and Pediatrics
UC Davis School of Medicine
Sacramento, CA

Amy K. Rosen, PhD
VA Research Career Scientist
VA Boston Healthcare System
Boston, MA

Jed Weissberg, MD
Senior Vice President, Quality and Care Delivery Excellence
Kaiser Permanente
Oakland, CA

APPENDIX C
GLOSSARY

Data element, critical: Quality performance measures are based on many individual items of information. The data elements are often patient-level information on individual patients (e.g., blood pressure, lab value, medication, surgical procedure, death). Testing at the data element level should include those elements that contribute most to the computed measure score, that is, account for identifying the greatest proportion of the target condition, event, or outcome being measured (numerator); the target population (denominator); population excluded (exclusions); and when applicable, risk factors with largest contribution to variability in outcome. Structural measures generally are based on organizational information rather than patient-level data.

Data element, quality: A quality data element is a single piece of information that is used in quality measures to describe part of the clinical care process, including both a clinical entity and its context of use (e.g., diagnosis, active).[14]

Electronic health record (EHR) (also electronic patient record, electronic medical record, or computerized patient record)**:** As defined by Healthcare Information Management and Systems Society (HIMSS), the electronic health record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports.[35]

Empirical evidence: Data or information resulting from studies and analyses of the data elements and/or scores for a *measure as specified*, unpublished or published.

Measure, eMeasure: As defined by Health Level Seven (HL7), an eMeasure is a health quality measure encoded in the Health Quality Measures Format (HQMF) format. The HQMF is a standard for representing a health quality measure as an electronic document. Through standardization of a measure’s structure, metadata, definitions, and logic, the HQMF provides for quality measure consistency and unambiguous interpretation.[36]

Measure, EHR: An EHR measure is a healthcare quality measure specified for use with electronic health records; it is composed of data elements from the quality data set (see below), including code lists and measure logic, and can be translated to computer-readable specifications.

Measure, quality (also quality performance measure)**:** Numeric quantification of healthcare quality for a designated healthcare provider, such as hospital, health plan, nursing home, clinician, etc.

Measure score: The numeric result that is computed by applying the measure specifications and scoring algorithm. The computed measure score represents an aggregation of all the appropriate patient-level data (e.g., proportion of patients who died, average lab value attained) for the entity being measured (e.g., hospital, health plan, home health agency, clinician, etc.). The measure specifications designate the entity that is being measured and to whom the measure score applies.

Measure testing: Empirical analysis to demonstrate the reliability and validity of the *measure as specified* including analysis of issues that pose threats to the validity of conclusions about quality of care such as exclusions, risk adjustment/stratification for outcome and resource use measures, methods to identify differences in performance, and comparability of data sources/methods.

Measure, untested: Measure without empirical evidence of both reliability and validity. Untested measures are only eligible for time-limited endorsement if the conditions for considering time-limited endorsement are met.

Quality Data Model (QDM, formerly QDS**):** Clinical data necessary to measure quality performance. The QDM framework contains three levels of information: standard elements, quality data elements, and data flow attributes. Standard elements (e.g., diagnosis) represent the atomic unit of data identified by a data element name, a code set, and a code list composed of one or more enumerated values. The quality data element includes the standard element plus quality data type or context (e.g., diagnosis active). Data flow attributes include source (originator), recorder, setting, and health record field.[14]

Reliability: Reliability refers to the repeatability or precision of measurement. Reliability of data elements refers to repeatability and reproducibility of the data elements for the same population in the same time period. Reliability of the measure score refers to the proportion of variation in the performance scores due to systematic differences across the measured entities (signal) in relation to random variation or noise.

Reliability testing: Empirical analysis of the *measure as specified* that demonstrate repeatability and reproducibility of the data elements in the same population in the same time period and/or the precision of the computed measure scores. Reliability testing focuses on random error in measurement and generally involves testing the agreement between repeated measurements of data elements (often referred to as inter-rater or inter-observer, which also applies to abstractors and coders) or the amount of error associated with the computed measure scores (signal vs. noise).

Reliability, threats: Some aspects of the measure specifications or the specific topic of measurement can affect reliability. Ambiguous measure specifications can result in unreliable measures. Small case volume or sample size, or rare events can affect the precision (reliability) of the measure score.

Validation: Process (testing) to determine if a measure has the property of validity. The term validation is often used in reference to the data elements and is another term for validity testing of data elements. Validation also is used in reference to statistical risk models where model performance metrics are compared between two different samples of data called the development and validation samples.

Validity: Validity refers to the correctness of measurement. Validity of data elements refers to the correctness of the data elements as compared to an authoritative source. Validity of the measure score refers to the correctness of conclusions about quality that can be made based on the measure scores (i.e., a higher score on a quality measure reflects higher quality).

Validity testing: Empirical analysis of the *measure as specified* that demonstrates that data are correct and/or conclusions about quality of care based on the computed measure score are correct. Validity testing focuses on systematic errors and bias. It involves testing agreement between the data elements obtained when implementing the measure as specified and data from another source of known accuracy. Validity of computed measure scores involves testing hypotheses of relationships between the computed measure scores as specified and other known measures of quality or conceptually related aspects of quality. A variety of approaches can provide some evidence for validity. The specific terms and definitions used for validity may vary by discipline, including face, content, construct, criterion, concurrent, predictive, convergent, or discriminant validity. Therefore, the proposed conceptual relationship and test should be described. The hypotheses and statistical analyses often are based on various correlations between measures or differences between groups known to vary in quality.

Validity, threats: In addition to unreliability, some aspects of measure specifications and data can affect the validity of conclusions about quality. Potential threats include patients excluded from measurement; differences in patient mix for outcome and resource use measures; measure scores generated with multiple data sources/methods; and systematic missing or “incorrect” data (unintentional or intentional).

APPENDIX D
MEASURE EVALUATION CRITERIA

Measure Evaluation Criteria (December 2009)

Conditions for Consideration
Four conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards:
A. The measure is in the public domain or an intellectual property agreement is signed.
B. The measure owner/steward verifies there is an identified responsible entity and process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every 3 years.
C. The intended use of the measure includes both public reporting and quality improvement.
D. The requested measure submission information is complete. Generally, measures should be fully developed and tested so that all the evaluation criteria have been addressed and information needed to evaluate the measure is provided. Measures that have not been tested are only potentially eligible for a time-limited endorsement and in that case, measure owners must verify that testing will be completed within 12 months of endorsement.

Criteria for Evaluation
If all four conditions for consideration are met, candidate measures are evaluated for their suitability based on four sets of standardized criteria: importance to measure and report, scientific acceptability of measure properties, usability, and feasibility. Not all acceptable measures will be strong—or equally strong—among each set of criteria. The assessment of each criterion is a matter of degree; however, all measures must be judged to have met the first criterion, importance to measure and report, in order to be evaluated against the remaining criteria.

1. Importance to measure and report: Extent to which the specific measure focus is important to making significant gains in health care quality (safety, timeliness, effectiveness, efficiency, equity, patient-centeredness) and improving health outcomes for a specific high impact aspect of healthcare where there is variation in or overall poor performance. ***Candidate measures must be judged to be important to measure and report*** *in order to be evaluated against the remaining criteria.*

1a. The measure focus addresses:

- a specific national health goal/priority identified by NQF's National Priorities Partners;
  OR
- a demonstrated high impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use (current and/or future), severity of illness, and patient/societal consequences of poor quality).

1b. Demonstration of quality problems and opportunity for improvement, i.e., data[1] demonstrating considerable variation, or overall poor performance, in the quality of care across providers and/or population groups (disparities in care).

1c. The measure focus is:

- an outcome (e.g., morbidity, mortality, function, health-related quality of life) that is relevant to, or associated with, a national health goal/priority, the condition, population, and/or care being addressed[2];
  OR
- if an intermediate outcome, process, structure, etc., there is **evidence**[3] that supports the specific measure focus as follows:
  - Intermediate outcome – evidence that the measured intermediate outcome (e.g., blood pressure, Hba1c) leads to improved health/avoidance of harm or cost/benefit.
  - Process – evidence that the measured clinical or administrative process leads to improved health/avoidance of harm and
    if the measure focus is on one step in a multi-step care process[4], it measures the step that has the greatest effect on improving the specified desired outcome(s).
  - Structure – evidence that the measured structure supports the consistent delivery of effective processes or access that lead to improved health/avoidance of harm or cost/benefit.
  - Patient experience – evidence that an association exists between the measure of patient experience of health care and the outcomes, values and preferences of individuals/ the public.
  - Access – evidence that an association exists between access to a health service and the outcomes of, or experience with, care.
  - Efficiency[5] – demonstration of an association between the measured resource use and level of performance with respect to one or more of the other five IOM aims of quality.

[1] Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, measure data from pilot testing or implementation. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

[2] Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, "never events" that are compared to zero are appropriate outcomes for public reporting and quality improvement.

[3] The strength of the body of evidence for the specific measure focus should be systematically assessed and rated (e.g., USPSTF grading system – grade definitions and methods). If the USPSTF grading system was not used, the grading system is explained including how it relates to the USPSTF grades or why it does not. However, evidence is not limited to quantitative studies and the best type of evidence depends upon the question being studied (e.g., randomized controlled trials appropriate for studying drug efficacy are not well suited for complex system changes). When qualitative studies are used, appropriate qualitative research criteria are used to judge the strength of the evidence.

[4] Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multi-step process, the step with the greatest effect on the desired outcome should be selected as the focus of measurement. For example, although assessment of immunization status and recommending immunization are necessary steps, they are not sufficient to achieve the desired impact on health status – patients must be vaccinated to achieve immunity. This does not preclude consideration of measures of preventive screening interventions where there is a strong link with desired outcomes (e.g., mammography) or measures for multiple care processes that affect a single outcome.

[5] Efficiency of care is a measurement construct of cost of care or resource utilization associated with a specified level of quality of care. It is a measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five IOM aims of quality. Efficiency might be thought of as a ratio, with quality as the numerator and cost as the denominator. As such, efficiency is directly proportional to quality, and inversely proportional to cost. (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; based on AQA Principles of Efficiency Measures).

If not important to measure and report, STOP.

2. Scientific acceptability of the measure properties: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

2a. The measure is well defined and precisely specified[6] so that it can be implemented consistently within and across organizations and allow for comparability. The required data elements are of high quality as defined by NQF's Health Information Technology Expert Panel (HITEP) [7] .

2b. Reliability testing[8] demonstrates the measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period.

2c. Validity testing[9] demonstrates that the measure reflects the quality of care provided, adequately distinguishing good and poor quality. If face validity is the only validity addressed, it is systematically assessed.

2d. Clinically necessary measure exclusions are identified and must be:

- supported by evidence[10] of sufficient frequency of occurrence so that results are distorted without the exclusion;

AND

- a clinically appropriate exception (e.g., contraindication) to eligibility for the measure focus[11];

AND

- precisely defined and specified:
  - if there is substantial variability in exclusions across providers, the measure is specified so that exclusions are computable and the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion);

[6] Measure specifications include the target population (e.g., denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (e.g., numerator), measurement time window, exclusions, risk adjustment, definitions, data elements, data source and instructions, sampling, scoring/computation.

[7] The HITEP criteria for high quality data include: a) data captured from an authoritative/accurate source; b) data are coded using recognized data standards; c) method of capturing data electronically fits the workflow of the authoritative source; d) data are available in EHRs; and e) data are auditable. NQF. *Health Information Technology Expert Panel Report: Recommended Common Data Types and Prioritized Performance Measures for Electronic Healthcare Information Systems*. Washington, DC: NQF; 2008.

[8] Examples of reliability testing include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing may address the data items or final measure score.

[9] Examples of validity testing include, but are not limited to: determining if measure scores adequately distinguish between providers known to have good or poor quality assessed by another valid method; correlation of measure scores with another valid indicator of quality for the specific topic; ability of measure scores to predict scores on some other related valid measure; content validity for multi-item scales/tests. Face validity is a subjective assessment by experts of whether the measure reflects the quality of care (e.g., whether the proportion of patients with BP < 140/90 is a marker of quality). If face validity is the only validity addressed, it is systematically assessed (e.g., ratings by relevant stakeholders) and the measure is judged to represent quality care for the specific topic and that the measure focus is the most important aspect of quality for the specific topic.

[10] Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers.

[11] Risk factors that influence outcomes should not be specified as exclusions.

– if patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that it strongly impacts performance on the measure and the measure must be specified so that the information about patient preference and the effect on the measure is transparent[12] (e.g., numerator category computed separately, denominator exclusion category computed separately).

2e. For outcome measures and other measures (e.g., resource use) when indicated:

- an evidence-based risk adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care[11,13]

OR

- rationale/data support no risk adjustment.

2f. Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful[14] differences in performance.

2g. If multiple data sources/methods are allowed, there is demonstration they produce comparable results.

2h. If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender);
OR
rationale/data justifies why stratification is not necessary or not feasible.

3. Usability: Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

3a. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audience(s) for both public reporting (e.g., focus group, cognitive testing) and informing quality improvement (e.g., quality improvement initiatives)[15]. An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.

[12] Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

[13] Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care such as race, socioeconomic status, gender (e.g., poorer treatment outcomes of African American men with prostate cancer, inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than adjusting out differences.

[14] With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74% v. 75%) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall poor performance may not demonstrate much variability across providers.

[15] Public reporting and quality improvement are not limited to provider-level measures – community and population measures also are relevant for reporting and improvement.

3b. The measure specifications are harmonized[16] with other measures, and are applicable to multiple levels and settings.

3c. Review of existing endorsed measures and measure sets demonstrates that the measure provides a distinctive or additive value to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of healthcare).

4. Feasibility: Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.

4a. For clinical measures, required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery.

4b. The required data elements are available in electronic sources. If the required data are not in existing electronic sources, a credible, near-term path to electronic collection by most providers is specified and clinical data elements are specified for transition to the electronic health record.

4c. Exclusions should not require additional data sources beyond what is required for scoring the measure (e.g., numerator and denominator) unless justified as supporting measure validity.

4d. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.

4e. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality[17], etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

If a measure meets the above criteria and there are competing measures (either endorsed measures, or other new submissions that also meet the criteria), compare measures on: Scientific acceptability of measure properties, Usability, and Feasibility to determine best-in-class.

5. Demonstration that the measure is superior to competing measures – new submissions and/or endorsed measures (e.g., is a more valid or efficient way to measure).

[16] Measure harmonization refers to the standardization of specifications for similar measures on the same topic (e.g., ***influenza immunization*** of patients in hospitals or nursing homes), or related measures for the same target population (e.g., eye exam and HbA1c for ***patients with diabetes***), or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are dictated by the evidence. The dimensions of harmonization can include numerator, denominator, exclusions, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.

[17] All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.

NOTES

1. McGlynn EA, Selecting common measures of quality and system performance, *Med Care*, 2003;41(1 Suppl):I39-I47.
2. McGlynn EA, Asch SM, Developing a clinical performance measure, *Am J Prev Med*, 1998;14(3 Suppl):14-21.
3. Rubin HR, Pronovost P, Diette GB, From a process of care to a measure: the development and testing of a quality indicator, *Int J Qual Health Care*, 2001;13(6):489-496.
4. Trochim WMK, Research methods knowledge base, *Web Center for Social Research Methods*, 2006. Available at: www.socialresearchmethods.net/kb/index.php. Last accessed May 2010.
5. Physician Consortium for Performance Improvement, *Measure Testing Protocol for Physician Consortium for Performance Improvement Performance Measures*, Chicago, IL: American Medical Association, 2007.
6. Bhattacharyya T, Freiberg AA, Mehta P, et al., Measuring the report card: the validity of pay-for-performance metrics in orthopedic surgery, *Health Aff (Millwood)*, 2009;28(2):526-532.
7. Schneider EC, Nadel MR, Zaslavsky AM, et al., Assessment of the scientific soundness of clinical performance measures: a field test of the National Committee for Quality Assurance's colorectal cancer screening measure, *Arch Intern Med*, 2008;168(8):876-882.
8. Moss PA, Can there be validity without reliability?, *Educational Researcher*, 1994;23(2):5-12.
9. Salvucci S, Walter E, Conley V, Fink S, Saba M, *Measurement Error Studies at the National Center for Education Statistics*, Washington, DC: U.S. Department of Education, 1997.
10. Adams JL, Mehrotra A, McGlynn EA, *Estimating Reliability and Misclassification in Physician Profiling*, Santa Monica, CA: RAND Corporation, 2010. Available at www.rand.org/pubs/technical_reports/TR863. Last accessed January 2011.
11. Reeves D, Campbell SM, Adams J, et al., Combining multiple indicators of clinical quality: an evaluation of different analytic approaches, *Med Care*, 2007;45(6):489-496.
12. Fitch K, Bernstein SJ, Aguilar MS, et al., *The RAND/UCLA Appropriateness Method User's Manual*, Santa Monica, CA: RAND Health, 2000. Available at www.rand.org/pubs/monograph_reports/MR1269/. Last accessed November 2010.
13. Spertus JA, Eagle KA, Krumholz HM, et al., American College of Cardiology and American Heart Association methodology for the selection and creation of performance measures for quantifying the quality of cardiovascular care, *Circulation*, 2005;111(13):1703-1712.
14. National Quality Forum, *Health Information Technology Expert Panel II - Health IT Enablement of Quality Measurement*, Washington, DC: NQF, 2009.
15. Baker DW, Persell SD, Thompson JA, et al., Automated review of electronic health records to assess quality of care for outpatients with heart failure, *Ann Intern Med*, 2007;146(4):270-277.
16. Persell SD, Wright JM, Thompson JA, et al., Assessing the validity of national quality measures for coronary artery disease using an electronic health record, *Arch Intern Med*, 2006;166(20):2272-2277.
17. Weiner M, Stump TE, Callahan CM, et al., Pursuing integration of performance measures into electronic medical records: beta-adrenergic receptor antagonist medications, *Qual Saf Health Care*, 2005;14(2):99-106.

18. Briggs JB, Kind EA, Awwad S, et al., *Performance Measures Using Electronic Health Records: Five Case Studies*, New York, NY: The Commonwealth Fund, 2008. Report No.: 1132, Available at www.commonwealthfund.org.
19. Adams JL, *The Reliability of Provider Profiling: A Tutorial*, Santa Monica, CA: RAND Corporation, 2009. Available at www.rand.org/pubs/technical_reports/TR653. Last accessed January 2011.
20. Kaplan SH, Griffith JL, Price LL, et al., Improving the reliability of physician performance assessment: identifying the "physician effect" on quality and creating composite measures, *Med Care*, 2009;47(4):378-387.
21. Austin PC, The reliability and validity of Bayesian measures for hospital profiling: a Monte Carlo assessment, *J Statist Plann Inference*, 2005;128(1):109-122.
22. Roebroeck ME, Harlaar J, Lankhorst GJ, The application of generalizability theory to reliability assessment: an illustration using isometric force measurements, *Phys Ther*, 1993;73(6):386-395.
23. Kahn JM, Iwashyna TJ, Accuracy of the discharge destination field in administrative data for identifying transfer to a long-term acute care hospital, *BMC Res Notes*, 2010;3:205.
24. Quan H, Parsons GA, Ghali WA, Validity of procedure codes in International Classification of Diseases, 9th revision, clinical modification administrative data, *Med Care*, 2004;42(8):801-809.
25. McGinn T, Wyer PC, Newman TB, et al., Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic), *CMAJ*, 2004;171(11):1369-1373.
26. Tooth LR, Ottenbacher KJ, The kappa statistic in rehabilitation research: an examination, *Arch Phys Med Rehabil*, 2004;85(8):1371-1376.
27. Viera AJ, Garrett JM, Understanding interobserver agreement: the kappa statistic, *Fam Med*, 2005;37(5):360-363.
28. Garson GD, Reliability analysis, *Statnotes: Topics in Multivariate Analysis*, January 2010. Available at: http://faculty.chass.ncsu.edu/garson/PA765/reliab.htm#rater. Last accessed November 2010.
29. Landis J, Koch G, The measurement of observer agreement for categorical data, *Biometrics*, 1977;33:159-174.
30. Zaslavsky AM, Statistical issues in reporting quality data: small samples and casemix variation, *Int J Qual Health Care*, 2001;13(6):481-488.
31. Nunnally J, Bernstein I, *Psychometric Theory*, 3rd ed., New York: McGraw-Hill, 1994.
32. Garson GD, Scales and standard measures, *Statnotes: Topics in Multivariate Analysis*, September 2008. Available at: http://faculty.chass.ncsu.edu/garson/PA765/standard.htm. Last accessed November 2010.
33. Cohen J, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., 1988.
34. Cohen J, A power primer, *Psychol Bull*, 1992;112(1):155-159.
35. Healthcare Information Management and Systems Society, Electronic Health Record, *HIMSS EHR Web Page*, 2010. Available at: www.himss.org/ASP/topics_ehr.asp. Last accessed November 2010.
36. Health Level Seven, 1.1 What is the HQMF, and what is an eMeasure?, *HL7 Version 3 Standard: Representation of the Health Quality Measures Format (eMeasure), Release 1 Last Ballot: Draft Standard for Trial Use - March 2010*, March 2010. Available at: www.hl7.org/v3ballot/html/domains/uvqm/UVQM.html#WhatisHQMF. Last accessed June 2010.