1

Introduction to the Measure of American Elections

Barry C. Burden and Charles Stewart III

How good are American elections?

Where would one start in answering this question?

Whenever this question is posed, it is common to answer it from the position of deeply held beliefs, but rarely from the position of a systematic analysis of facts. These beliefs might arise from partisanship: a good election is one that my favored candidate wins. These beliefs might be chauvinistic: a good election is one run according to the rules of my community.

Rarely are these beliefs rooted in hard facts.

When facts intervene, they rarely are presented in a systematic fashion. Opinions about levels of voter fraud might be attributable to a viral YouTube video. Concerns about the effects of a new voter identification law might be informed by a reporter's interview with an activist who is eager to share stories about how voters she has talked with will be disenfranchised on Election Day. Satisfaction with a new electronic voting machine may be illustrated by a picture of a smiling citizen coming out of the precinct with an "I Voted" sticker stuck to her lapel. Disdain about the ability of local governments to run elections might follow from a newspaper article detailing yet another season of long lines when waiting to vote in Florida (or South Carolina or Maryland or …). At its worst, this approach is evaluation by anecdote.

Consider instead how the question about the quality of American elections would be framed if first we asked about other policy domains: "How good are America's prisons?" or "How good are America's schools?" or "How good is America's health care system?" Some people surely would respond based on fact-free beliefs; others would respond with a random story about the experience that one's cousin had with one of these institutions. However, it would not be difficult to discover that in

1

2007 (the most recent year for which data are available), 15.5 percent of all parolees were reincarcerated, that Connecticut had the highest reincarceration rate (29.9 percent), and that Maine had the lowest (0 percent).[1] Nor would it be difficult to find out that Alaska's fourth graders ranked last among the fifty states in the reading portion of the 2011 National Assessment of Educational Progress and that Massachusetts ranked first; that the gap between girls and boys was greatest in Hawaii and smallest in Texas; and that the gap between whites and blacks was greatest in Connecticut and smallest in North Dakota.[2] A brief Internet search would reveal that in 2008, the infant mortality rate among the fifty states ranged from 3.87 per 100,000 live births in New Hampshire to 9.95 per 100,000 in Mississippi (Mathews and MacDorman 2012, 6).

In other words, an obvious way to begin addressing questions about the state of public policy in these other important areas would be to draw on a large body of data about the performance of these institutions and policy systems.

None of the statistics just referenced is the be-all and end-all of the question about how well the prison systems, schools, and health care systems work in the states. The point is that in each of these policy domains, significant effort is poured into defining measures of policy input and output consistently across states, multiple measures of system performance are regularly reported through a federal agency, and entire professions have grown up to analyze these data. Despite the fact that answers to policy questions about criminal justice, education, and health care are legitimately informed by ideology and deeply held beliefs, even committed ideologues typically ground their appeals in statistics when they argue about policy; some will even be convinced they are wrong if the facts are against them. The data provide a common starting point.

This returns us to the original question: How good are American elections? If an American wanted to argue this question based on facts, he would most likely go to the turnout statistics and discover that in 2012, 58.2 percent of eligible Americans voted, ranging from 44.2 percent in Hawaii to 75.7 percent in Minnesota.[3] Compared with other nations, U.S. turnout was in the lower half of 112 countries

[1] Probation and Parole in the United States, 2007 – Statistical Tables, http://bjs.ojp.usdoj.gov/index.cfm?ty=pbdetail&iid=1099.
[2] The Nation's Report Card – National Assessment of Educational Progress, http://nces.ed.gov/nationsreportcard/naepdata/.
[3] United States Election Project, http://elections.gmu.edu/voter_turnout.htm.

with presidential elections in recent years – ranked seventy-fourth (right between the Democratic Republic of Congo and Romania), above the European nations of Poland, Serbia, Portugal, Slovakia, Ireland, Austria, Macedonia, and Lithuania, but below the nations of Moldova, the Russian Federation, Montenegro, Finland, Ukraine, Iceland, France, Kazakhstan, Armenia, Cyprus, Tajikistan, Belarus, Turkmenistan, and Uzbekistan.[4] An American with a bit more perseverance might also discover that when asked in a national survey in 2012 whether they were confident that their votes were counted as cast, 63 percent stated they were "very confident," ranging from 80 percent in Vermont to 54 percent in Washington.[5]

An American who was interested in understanding in a more nuanced sense how well elections are run in this country would *not* find much in the way of defining consistent measures of election administration input and output across states. Nor would he or she find much effort at reporting such statistics even *within* states, much less a profession devoted to the proposition that elections would function better if we understood systematically the facts associated with election administration. Indeed, this American would find that states differ in how they even define critical aspects of election administration, including such fundamental measures as turnout. Sometimes even a state's chief election officer lacks the authority to require local election boards to report to him or her basic statistics about election administration beyond the bare facts of the election returns themselves. This American would also discover that the National Association of Secretaries of State (NASS), which is the organization of top state constitutional officers who most often are ultimately responsible for the conduct of elections, has led a campaign to abolish the U.S. Election Assistance Commission (EAC), the only federal agency that gathers statistics about election administration nationwide.

In other words, while there are scientific professions devoted to the study of corrections, education, public health, transportation, and many other critical functions of state and local government, there is no scientific profession devoted to the study of election administration.

This book is part of an effort to change that.

[4] International Institute for Democracy and Electoral Assistance, Voter Turnout Database, http://www.idea.int/vt/viewdata.cfm.
[5] See "2012 Survey of the Performance of American Elections" (Stewart 2013).

In particular, the chapters in this volume are devoted to the study of ten areas of election administration through the lens of hard data and social science. The topics cover the waterfront in the field of election administration and policy, ranging from the registration of voters to the counting of votes. None of these chapters is the final word in any of these areas. They are something more important: in many cases, they are the *first word* in starting a conversation about the systematic analysis of election administration and policy in America.

<div align="center">

FIRST THINGS FIRST:

ESTABLISHING THE VALIDITY AND RELIABILITY OF AMERICAN
ELECTION ADMINISTRATION

</div>

In the areas of public policy that aspire to be data driven, great attention is paid to the validity and reliability of the key measures employed in the field.

Validity may be defined as the degree to which a measure actually describes the underlying concept it claims to measure, rather than something else. In other words, it refers to how well an observable quantity describes an unobservable theoretical construct. For instance, we cannot observe someone's intelligence directly, but we can observe how well she or he performs on an IQ test. A robust literature and scholarly debate has grown up around the question of how valid IQ tests are as an indicator of (unobservable) intelligence.[6]

An example of the application of the concept of validity to elections is in the area of absentee ballots. For example, a valid measure of absentee ballot usage should track actual usage, even if it sometimes misses the mark somewhat. Similarly, the measure of line lengths is highly valid if it approximates the actual waiting times, rather than being consistently too high or too low.

There are many methods used to establish validity, including such things as relating a measure to other variables known to affect or be affected by it. This is often an interactive process of moving back and forth between measures and their actual scores (Adcock and Collier 2001).[7]

---

[6] Within political science, the classic statement on validity and reliability is Carmines and Zeller (1979).

[7] This sort of refinement is discussed in some of the chapters that follow. Here we mention two examples. The absentee ballot rejection rate is an important measure that could be computed

In contrast, *reliability* refers to the degree to which a measuring procedure yields the same results when the procedure is repeated. A highly reliable measure produces similar results when applied multiple times in the same setting. For example, a reliable measure of a state's absentee ballot usage will yield a similar conclusion even if different staff members provide the data. Likewise, a reliable measure of polling place lines based on surveys of voters will show consistent times if the survey is repeated. Although low reliability does not affect accuracy of a measure, it does affect its precision. That is, while the answer might be right on average, the high degree of variability makes it difficult to discern the signal amid the noise. One way to increase reliability is to bring more data to the table by adding observations or combining measures into a summary index.

In psychometrics a great deal of attention is paid to whether respondents give the same answers when given the exact same battery of questions on different days. Similarly, a diagnostic medical test that gives the same results when repeated on the same individual in rapid succession is said to be reliable.

Measures of validity and reliability are generally measures of *degree* and not absolutes. It is well understood that reliable measures may fluctuate from one moment to the next, as when the results of an aptitude test vary depending on whether a student is well rested, under stress, or hungry. Thus, the standard is rarely perfection when it comes to assessing validity and reliability, but the closer we can come to perfection, the better.

On a scale where 1.0 indicates perfect reliability, most of the state measures used in the chapters that follow correlate between the 2008 and 2010 elections at levels between .7 and .9. Where the correlations are lower, it is sometimes a sign of poor reliability stemming from a small sample size. Aggregating over multiple elections will remedy this problem. In other cases, it is a sign that what the measure is capturing has changed between the presidential election and the midterm election. Several of the Uniformed and Overseas Citizens Absentee Voting Act (UOCAVA) and absentee ballot measures are of this variety.

as a share of absentee ballots cast or of all ballots cast. Which is used depends on both theory about the denominator of interest and also empirical information about which offers greater discrimination across states. Voter confidence is also of keen interest. Survey questions often ask whether the respondent is "very confident," "somewhat confident," "not too confident," or "not at all confident." Where to "cut" these four categories to create state-by-state percentages is partially a result of trial and error to find the most meaningful division.

In some fields, the reliability and validity of performance mea-
sures are based on a deep body of scientific research, stretching across
decades and thousands of researchers, supported by billions of dollars
of basic research. Consider, for instance, a study to assess the quality of
medical care received by Medicare recipients in each of the fifty states,
which was published in the *Journal of the American Medical Association*
(*JAMA*) in 2000 (Jencks et al. 2000). To assess the quality of medical
care received by Medicare patients, this study examined thousands of
randomly chosen patient records, drawn from patients in every state in
the nation. The basic question asked of these records was whether the
patient in question was treated for his or her conditions in a way that
reflected the consensus of appropriate care in a particular field. The
overall index of patient care was based on rating twenty-four indicators
of patient care such as these; the conditions covered by these indicators
affected 85 percent of Medicare beneficiaries.

The validity of the measures chosen by the researchers in the *JAMA*
study had already been established in the medical community, based
on hundreds of studies of the relationship between the care patients
received and how well they fared after their treatments. For instance,
these previous studies had established that one valid measure of the
quality of care of Medicare recipients was what percentage of those
admitted to hospitals because of acute myocardial infarction (better
known as a heart attack) later received prescriptions for aspirin upon
discharge.

The reliability of such an indicator would be more difficult to mea-
sure, because it is based on the assessment of patient records. Although
there may be *general* guidelines for care of patients, there also may be
well-established contraindications for applying that regimen to a *par-
ticular* patient – such as if the patient is known to be violently allergic
to the drug of choice. Here, judgment must be applied to the coding
of patient records in the making of the index. This is where reliability
comes in.

Two physicians, and certainly two research assistants, reading the
same patient medical records might come to different conclusions
about whether a particular patient is a candidate to receive treatment
consistent with the conventional standards of care. If they each read
the same record and come to different conclusions, we would infer that
this method of coding whether patients received good hospital care
based on their records is unreliable. Conversely, if the coders frequently
agreed, we would regard this measure as more reliable.

The reliabilities of the measures used in the *JAMA* study were assessed precisely according to this logic. Reliability was established by testing to see whether two independent coders came to the same conclusions about whether to include or exclude a particular patient record from the study. Part of the study's report included measures of inter-rater reliability for each of the indicators.

The effort just described to compare the quality of hospital care received by participants in the Medicare program across the fifty states rested on decades of medical research that consumes billions of dollars in taxes and private foundation support each year. The study of election administration barely has thousands of dollars devoted to it each year, so the reliability and validity of any measures of election administration quality that we consider must be viewed as provisional.

To continue with the medical analogy, election administration is similar to the earliest days of public health studies, when the only measure of health care quality was the mortality rate – the number of deaths divided by the number of people. This was a crude measure, but it was powerful. It could establish, for instance, that modern sewer systems improved the health of city dwellers, leading to greater attention on public works projects aimed at improving the quality of life of everyone and, ultimately, promoting economic growth.

As the raw mortality rate was used increasingly to compare health care outcomes across geographical units, hospitals, and demographic populations, it became clear that the validity of the crude mortality rate could be improved by adjusting based on the risk of dying, which might be quite independent of the quality of health care (or other public health factors). For instance, the mortality rate of a city with a high fraction of elderly residents would be higher than that of a city with few elderly residents, even if the sanitation facilities were equivalent in the two places. Thus, as the field of public health has advanced, the mortality measure has been improved by adjusting for risk factors such as age. In doing so, the new risk-adjusted mortality statistics have become even more valid measures of the underlying health of a given population.

Election administration is still in the "raw mortality rate" phase of measuring outcomes in its policy domain. Few measures of performance within election administration are widely understood and accepted.

Indeed, it could be argued that there is only one widely known measure of performance in this area: the turnout rate, which shares many of the advantages and disadvantages of the raw mortality rate. On the plus side, the turnout rate is intuitively understood and easily calculated.

Most of the country understood, for instance, that it was bad when only 25 percent of Mississippi's voting-age population (VAP) turned out to vote in the 1960 presidential election. The country also understood that when the turnout rate in Mississippi rose to the 60 percent level in 2012, this was a valid indication that the state's electoral system had improved over the intervening half-century.

On the negative side, the raw turnout rate does not take into account the "luck" or "skill" that might be involved. For instance, what do we make of the fact that the turnout rates of Hawaii and Texas in the 2008 presidential election were 44 percent and 50 percent, respectively, despite levels of educational attainment – a factor that strongly predicts whether an individual will vote – being much higher in Hawaii (90 percent high school graduates) than in Texas (81 percent high school graduates)?[8] Would we consider, on a risk-adjusted basis, that turnout in Texas was actually *much* higher than in Hawaii, because efforts to get voters to the polls in the Lone Star State must battle against lower educational levels than those in Hawaii? Or what about the comparison of Minnesota and Mississippi? In raw turnout terms, it is no contest. The Minnesota turnout rate was 75 percent in 2008, whereas Mississippi's was 61 percent – a fourteen-point deficit for Mississippi. But if we account for different levels of educational attainment using a simple technique that relies on linear regression, the tables are turned, with Minnesota suffering a five-point turnout deficit in "risk-adjusted turnout" in comparison with Mississippi.[9]

As far as we are aware, no report of turnout rates that compares the states has ever reported turnout rates adjusted for "risk factors" such as education so that the effects of policy or other systemic factors on turnout can be better understood. The best studies to this point are multivariate analyses that attempt to control for "luck" factors, such as the demographics of the state's electorate or efforts by campaigns to turn out voters. At least turnout levels of the states are periodically reported to the public and discussed. The same cannot be said of a

[8] Educational attainment levels are taken from the 2011 American Community Survey (three-year series).

[9] The method of risk adjusting here is based on a simple regression method, in which we regress the turnout rate of each state on the percentage of residents twenty-five years of age and up with a high school diploma and then generate the residuals of the regression. In this case, the Minnesota residual is +5.8 percentage points, and the Mississippi residual is +10.4 percentage points. The regression equation in this case was Turnout Rate = –84.8 + 1.68 H.S. Graduate Rate, with standard errors of 15.9 and 0.18, respectively. The R-squared statistic was .63, with 51 observations. The regression was weighted by turnout in each state.

long list of other measures of election administration that could easily
be constructed from official sources, such as the rejection rate of provi-
sional ballots or the nonreturn rates of absentee ballots.

The ideal measure is one that is both reliable and valid. That is, it has
a high level of precision and a high level of accuracy. It is possible for a
measure to do well on one dimension but poorly on the other, or poorly
on both. A digital watch is highly precise, but it can be inaccurate if set
to the wrong time; likewise, a mechanical watch without a second hand
might be highly accurate while being far from precise. A key purpose
of this volume is to evaluate a variety of election-related measures on
these two dimensions. It is important to note that for the purposes of
evaluating election performance, an invalid measure is not necessarily
useless. If a measure of provisional ballot usage, for example, systemati-
cally underestimated the actual use across the states, it would be less
valid but could still be used effectively to rank order states. In contrast,
a valid measure with low reliability would produce problematic rank-
ings because of the excessive noise in the indicator, but it could still be
used to describe how states perform on average.

WHERE ARE THE DATA?

SOURCES OF DATA FOR THE ASSESSMENT OF AMERICAN ELECTIONS

This volume begins the process of assessing how elections are con-
ducted in America by identifying a manageable set of candidate indi-
cators, subjecting them to scrutiny, and examining them for what
they tell us about elections in America. If we agree that data-driven
scrutiny of American elections is to be commended, we need data.
Luckily there are plenty of data out there, oftentimes hiding in plain
view. The remaining chapters of this book rely on these data sources
to quite a detailed level. Here we introduce the reader to the most
important.[10]

There are four major sources of data available for the assessment of
American elections:

1. The Voting and Registration Supplement of the Current Population
   Survey (CPS), conducted by the U.S. Census Bureau.

---

[10] A more detailed examination of many of these data sources may be found in Pew Center on
the States (2012).

2. The Election Administration and Voting Survey (EAVS), conducted by the U.S. EAC.
3. State and local election board records.
4. Academic and commercial survey research.

The Voting and Registration Supplement

Ever since the 1960s, the U.S. Census Bureau has conducted a survey every other November about voting in the most recent federal general election, as a supplement to the Census Bureau's monthly CPS. The primary purpose of the CPS is to help determine the unemployment rate and other economic statistics at the state level. The biennial survey that studies voting is called the Voting and Registration Supplement (VRS). The main CPS contains a treasure trove of information about the participants in the survey, including information about their ethnicity, education, housing, income, and (since 2010) disability status. Thus, it is possible to study the relationship between important demographic characteristics of American adults and the likelihood they will be registered and vote, and even the means used to register and to vote. The overall survey sample is quite large. In 2010, for instance, nearly 80,000 respondents were asked whether they voted in the most recent federal election, ranging from 578 in New Mexico to 5,862 in California.

Compared with other survey research that focuses on election behavior, the VRS is actually quite limited in what it asks about voting – some would say "focused." It does not ask respondents whom they voted for. (This being a survey conducted by the federal government, it is easy to understand why.) Nor does it ask other questions that political scientists studying participation might want to consider, such as interest in politics, knowledge of politics, or stances on important issues. It essentially only asks respondents whether they voted and if they are registered; if they report they did not vote or were not registered, they are asked why not. Those who say they voted are asked what mode they used to vote (in person on Election Day, in person before Election Day, or via the mail). Finally, they are asked how long they have lived at their current address and how they registered to vote.[11]

[11] This description of items is based on the 2010 VRS. While the items change infrequently, there is some fluctuation in questions from time to time. For instance, after the controversy that arose in 1980 over Jimmy Carter conceding defeat in the 1980 presidential election before the polls had closed on the West Coast, the VRS asked respondents that year what time of day they voted.