“Use It or Lose It” Professional Judgment:

Educational Evaluation and Bayesian Reasoning

Sherman Dorn

University of South Florida

Author correspondence: sdorn@coedu.usf.edu

Suggested citation:

Dorn, S. (2009). *“Use it or lose it” professional judgment. Educational evaluation and Bayesian reasoning*. Working paper published through the Social Science Research Network (papers.ssrn.com). Accessed [date], from http://ssrn.com/abstract=1461508.

My thanks to David Figlio, Doug Harris, Jennifer Imazeki, and Jeffrey Kromrey for ideas and comments on earlier versions. Any errors of interpretation or facts remain the author’s.

Running Head: USE IT OR LOSE IT

Abstract

This paper presents a Bayesian framework for evaluative classification. Current education policy debates center on arguments about whether and how to use student test score data in school and personnel evaluation. Proponents of such use argue that refusing to use data violates both the public's need to hold schools accountable when they use taxpayer dollars and students' right to educational opportunities. Opponents of formulaic use of test-score data argue that most standardized test data is susceptible to fatal technical flaws, is a partial picture of student achievement, and leads to behavior that corrupts the measures. A Bayesian perspective on summative ordinal classification is a possible framework for combining quantitative outcome data for students with the qualitative types of evaluation that critics of high-stakes testing advocate. This paper describes the key characteristics of a Bayesian perspective on classification, describes a method to translate a naïve Bayesian classifier into a point-based system for evaluation, and draws conclusions from the comparison on the construction of algorithmic (including point-based) systems that could capture the political and practical benefits of a Bayesian approach. The most important practical conclusion is that point-based systems with fixed components and weights cannot capture the dynamic and political benefits of a reciprocal relationship between professional judgment and quantitative student outcome data.

"Use It or Lose It" Professional Judgment: Educational Evaluation and Bayesian Reasoning

On July 24, 2009, President Barack Obama and Secretary of Education Arne Duncan announced draft regulations for state applications for "Race to the Top" funds appropriated by Congress in early 2009 (Branigin, 2009). Among the draft requirements for state applicants was the elimination of so-called legislative and regulatory "firewalls" between student-outcome data and teacher records. The assertive rhetoric from Obama administration appointees clearly implies that the current administration will push states to use student outcome data in teacher evaluation. While most of the public discussion of and controversy surrounding such data use has focused on performance-pay policies (e.g., Azordegan, Byrnett, Campbell, Greenman, & Coulter, 2005; Behrstock & Akerstrom, 2008; Max & Koppich, 2007), the most consequential potential use of student outcome data is for questions of employment—whether and to what extent test-score and other outcome data will influence the ordinary evaluation and continuation of teacher employment, tenure, and intervention efforts (Baratz-Snowden, 2009; Weisberg, Sexton, Mulhern, & Keeling, 2009).

In the past decade, arguments about test-score use have focused on the No Child Left Behind Act's cruder mechanisms for labeling schools, either arguments in favor of formulaic triggers as essential public-policy tools to equalize opportunity or arguments that such triggers are inherently corrupting (e.g., Dorn, 2007; Nichols & Berliner, 2007). Similar arguments inevitably surround the use of test-score and other student outcome data for personnel evaluation purposes, but with additional issues: the attribution of outcomes to single teachers, the omission of many educators by default when assessments only exist for a small part of the curriculum, and a recent technical literature emphasizing the difficulty of identifying anything more than a small portion of teachers as outliers (as either effective or ineffective) through test-score data (e.g., Baratz-Snowden, 2009; Lockwood, Louis, & McCaffrey, 2002).

While Duncan has made clear that his intent is not to force teacher evaluation decisions to revolve entirely around test scores, the current administration's position is that student outcomes need to figure into evaluation. But how that is done is an open question. In Florida, the statutory authorization for the Merit Award Program option for school districts requires that student outcome data "shall be weighted at not less than 60 percent of the overall evaluation" (Florida Statutes 1012.225(3)(c)). The Florida legislature mandated one of several options to use in combining quantitative and qualitative judgments of teacher effectiveness, the point system. While there are variations that meet the statutory language, almost any real-world implementation meeting the spirit of the law would almost all be linear combinations of different subscores (also see Max, 2007).

This type of algorithm for combining qualitative professional judgments and student outcome data is not the only option for using student outcome data. In performance-pay, for example, there exist a small number of programs that include student performance as one pathway through which teachers may seek pay increases, as in Minneapolis or Denver (Azordegan et al., 2005; Potemski & Rowland, 2009).[1] Baratz-Snowden (2009) argued that student outcomes should be part of evaluation, but without specifying how:

> Standardized test scores can play a role in presenting evidence of learning, but using standardized test scores as the sole or predominant measure of

[1] Also see the Denver ProComp website at http://denverprocomp.org.

> achievement is unwarranted and unwise given the inadequacy of such tests to capture the complexities and breadth of student learning and the limitations of current value-added methodologies. Nonetheless, it is absolutely essential that teachers present evidence of student learning—through test results and other material—as part of the tenure system if it is to be credible. Calling upon experienced teachers to help develop the multiple sources of such evidence is essential in redesigning the tenure system. (p. 28)

The evaluation systems she highlighted—in Toledo and Minneapolis's local public school systems and in the Los Angeles Green Dot union contract—are different variations of a holistic or portfolio system of teacher evaluation, including requirements for documenting both the process and outcomes of professional development.

Such a holistic system may well be the outcome of proactive collaboration between teacher union locals and local school boards, but the history of performance-pay plans suggests that some states will attempt to impose the type of algorithmic requirements for evaluations that Florida's legislature has created in its performance-pay statute. In states without collective bargaining or with more legal or practical authority for legislatures, local collective bargaining may be less important than the political environment at the state level. Legislators who distrust school districts are going to be less likely to accept holistic evaluation reviews than district-level management with a history of collaborative relationships with unions.

The differences between holistic and algorithmic use of student outcome data include at least two dimensions of the continuing debate over high-stakes testing policies. One dimension is the technical adequacy of existing assessments. Advocates of an algorithmic approach are likely to argue that current assessments are not perfect but are a sufficient basis on which to make decisions. The same recognized flaws of existing assessments will probably be the focus of critics of an algorithmic approach (whether or not the critics would accept even a holistic teacher evaluation system that uses student outcome data). The critics will continue to argue that tests assess only a small portion of the formal curriculum and student performance, and that their use will corrupt the measures and teaching practices. Behind the debates about the technical flaws or adequacy of existing assessments, however, there is another dimension of the discussion, and that is around trust of professional judgment. The arguments of Weisberg et al. (2009) feed into the historical dynamic of accountability politics (Dorn, 2007): policymakers and many citizens distrust either the capacity or willingness of educators to make appropriate judgments about school practices and teacher performance. Reciprocating this lack of trust, many teachers believe that state legislators and advocates of high-stakes testing use testing as a tool with which to attack public schools and teachers.

Unless addressing trust and mistrust is central to the design of teacher evaluation systems, the evaluation policies that develop in response to criticism of current practices are likely to be unsatisfactory to the two sides of the debate over test-score use. Many school critics are wary of evaluation policies that leave open the possibility of evaluation systems that never identify weak teachers, and an algorithmic approach such as the one mandated for Florida’s Merit Award Program is likely to appeal to such critics. But an algorithmic approach will be unsatisfactory to those who distrust the use of tests to drive decisionmaking in schools.

Bayesian Perspectives on Classification

Resolving the trust problem in teacher evaluation requires stepping back to ask what we are seeking: sound decisions about whether teachers should continue without intervention, should be given additional professional assistance, or should leave the field. Making those decisions with confidence should be goal of a teacher evaluation system, decisions that are far more likely to be right than wrong. There is one algorithmic approach that could be promising, or at least a foundation for thinking about how to combine qualitative professional judgment and quantitative data in ways that focus on making critical personnel decisions, give significant weight to professional judgment when it is made, and leaves a safety valve for decisionmaking when supervisors (and peers, in peer-review systems) are unwilling to make hard decisions about teachers. One can use Bayesian reasoning to understand evaluative classification as a process of judgments reshaped with data. This section describes Bayes' Theorem on the calculation of conditional probabilities, a possible translation of that use in personnel and program evaluation, and some of the general political and technical issues involved in translating Bayes' Theorem into evaluation use.

Bayes' Theorem and Conditional Probability

The standard presentation of Bayes' theorem centers on the conditional probability of A given data x, or P(A|x),

$$P(A|x) = \frac{P(A)P(x|A)}{P(x)} \tag{1}$$

where P(A) is the general probability of A, P(x) is the general probability of x, and P(x|A) is the conditional probability of observing x given A (also called the likelihood of observing *x* given A).[2] Equation (1) is the most common form of Bayes' Theorem, and it captures the relationship among four components of conditional probability. With complete information, one can see Bayes' Theorem in action. For example, if one examines the 1785 public elementary schools in Florida which received a letter grade from the state in June 2009, what is the probability of receiving a letter grade of "A" if the percentage of tested students meeting the state's standards on the reading test was exactly 50%?[3] Here, almost 71% of all public elementary schools in Florida received an "A" from the state, 0.08% of "A"-labeled schools had exactly 50% of all participating students meeting the state's standards on reading tests, and 0.56% of all schools had 50% of students meeting state standards, or

$$P(A|x) = \frac{P(A)P(x|A)}{P(x)} = \frac{0.706*0.0008}{.0056} = 0.10.$$

This claimed is validated by inspection of the records: 10 public elementary schools in Florida had 50% of all students meeting state standards in the spring 2009 tests, and one such school (or 10%) received an "A" from the state in summer 2009. (Out of all Florida elementary schools receiving an "A," Liberty City Elementary School in Miami had the lowest proportion of students at or above the state reading test cut score in spring exams.) A similar exercise with the set of all public elementary schools having 50% or more students passing state standards in 2009 (1723 schools, a set that contained all 1261 elementary schools receiving an "A" from the state in 2009) will show that 73.2% of such schools received a letter grade of "A."

[2] The appendix contains a more technical discussion of Bayesian reasoning, the naïve Bayesian classifier, and a translation to an additive point system.

[3] For letter grades assigned Florida's public schools, see http://schoolgrades.fldoe.org.

With complete data, Bayes' theorem is an accounting exercise. But an accounting exercise is not the value of Bayes' theorem. The general value of conditional probability is the ability to reason consistently about incomplete information. The evaluation of medical test results is the most common example of this use (perhaps because test results are often evaluated with the wrong perspective). If someone tests positive for a rare condition—for example, if the probability of having the condition is 1 in 10,000—even a highly accurate test can generally be wrong, even where 95% of those with the condition have a positive test, and only 5% of those without the condition test positive. Here, we break down P(x) into the sum of the probability of testing positive for those with the disease (A) and the probability of those testing positive without the disease (~A):

$$P(A|x) = \frac{P(A)P(x|A)}{P(A)P(x|A)+P(\sim A)P(x|\sim A)} = \frac{0.0001*0.95}{0.0001*0.95+0.9999*.05} = \frac{0.000095}{0.000095+0.049995} = \frac{0.000095}{0.05009} = 0.19\%.$$

The result is counterintuitive to many: a test with 95% accuracy in two dimensions is going to be wrong for the vast majority of positive results from a population where the risk is extremely low. Because the prevalence of a condition can dominate the value of a test result, repeating tests (or testing split samples) is important to provide confidence about the interpretation of positive test results for rare conditions.

Bayes and Inductive Reasoning

A minority of statisticians and a number of philosophers of science push a Bayesian approach in a different direction; in one Bayesian perspective, P(A) could be the general probability of A, but it can also be the judgment of the probability of A before gathering data, or the prior probability. In this view, P(A|x) is the posterior probability of A after gathering data x. This approach combines a prior judgment of A (which could be based on qualitative judgments) with data collection and analysis. In a Bayesian perspective, the data updates (or bumps) one's prior judgment. Bayesian advocates argue that this is consistent with the scientific method (Howson & Urbach, 2005). Those skeptical of a Bayesian approach with a subjective prior often argue that the inclusion of subjective judgment in a prior probability is not objective; subjective Bayesians often respond that the priors always exist, and a subjective Bayesian approach merely reveals those choices in an explicit fashion.

There is little literature attempting to apply a Bayesian approach to program or personnel evaluation in education, either the ordinary meaning of conditional probability or the subjectivist Bayesian approach.[4] The social-science field with experience in applying a Bayesian reasoning to conditional probability is in law, where there is a small literature on discussing statistics with juries (e.g., Kaye, 1999; Lindsey, Hertwig, & Gigerenzer, 2003). Some litigators have an incentive to avoid juries' inappropriately applying conditional probability in a manner known in legal circles as the prosecutor's fallacy—confusing the probability of testing positive given a hypothesis with the probability of the hypothesis being true given a positive test (e.g., Fenton & Neil, 2000).

Bayesian Reasoning and Evaluation Policy

The current policy debate over teacher evaluation provides an important reason to consider the use of a Bayesian approach, an approach with both practical and political benefits. For these purposes, the most important characteristic of the standard equation

[4] See Wood (1972) for a book review of Bayesian arguments at a Phi Delta Kappan symposium, with subjective Bayesian reasoning apparently considered more charming than practical by Wood.

for a posterior probability is the relationship between the prior probability and the likelihood: a forceful statement of prior probability is bumped less by any given data than a weaker statement of prior probability.[5] For classification purposes, one would compare the probability of being in two different groups, or the odds. For example, consider an evaluation system that uses principal or peer judgment, and a teacher where both the school principal and peers think the teacher could use intervention but are not entirely certain. If the professional judgment before gathering additional data is that the teacher is somewhat more likely to need intervention than not—odds of 3:2, or a professional judgment that 60% of teachers in similar situations and with similar information available to administrators and teachers would need remediation—how could additional data update that professional judgment? The key term is a likelihood ratio—for a choice between one decision and another (for example, deciding whether a teacher needs help with instruction), the ratio of the likelihood of seeing a data pattern under one hypothesis (for example, intervening with a teacher) against the likelihood of seeing the *same* data pattern under the competing hypothesis (not intervening). After gathering additional data—and again assuming that the data is professionally relevant and the relevant likelihood functions are salient—assume that the principal and peers discover that 6% of teachers judged as needing remediation produce the data gathered and that 1% of teachers judged as *not* needing remediation produce the data gathered. The likelihood ratio is 6:1, and the posterior (after-data-gathering) odds of needing intervention then become 9:1, or a 90% posterior probability of needing remediation. But the data can also bump the prior judgment in the other direction. If 6% of teachers

[5] In the long run, data will dominate both a frequentist and a Bayesian's estimation of relevant quantities. But we generally do not live in an asymptotic world, especially with regard to personnel evaluation.

judged as needing remediation produce the data gathered but 8% of teachers judged as *not* needing remediation also produce the data gathered, the likelihood ratio is 6:8 (or 3:4), and the posterior (after-data-gathering) odds of needing intervention then become 9:8, or an only slightly greater than even odds (approximately 53% probability) of needing intervention.

In the hypothetical cases described above, data can help one update the prior judgment, and data can bump the prior judgment in different directions. But there is another important characteristic of posterior odds: the forcefulness of the prior odds also shape the posterior judgment. A forceful statement of prior relative odds (e.g., a prior judgment that the teacher is twice as likely to need intervention as not) would be bumped less by any given data than a weaker statement of prior relative odds (e.g., a prior judgment that the teacher is equally likely to need intervention as not). If a school's culture is one where the principal and peers err on the side of nonintervention in the case of a weak teacher, then the likelihood ratio would dominate the posterior odds. If a principal and peers make forceful judgments about teachers, the data are going to be less influential.[6]

The reciprocal relationship between the influence of prior judgments and the influence of data-generated likelihood ratios could well be a practical and political strength rather than a liability, including and perhaps especially for those skeptical of subjective judgments of teacher effectiveness. A system with a Bayesian rationale for combining professional judgment with quantitative data can encourage professional

[6] As Howson and Urbach (2005) and many others point out, a large amount of data will dominate strong prior probabilities. The practical issue here is balancing professional judgment and student outcome data, and in many cases the data from a single year will not dominate a prior supplied by a professional evaluative rating.

judgment by making the judgment more influential where administrators (and peers in peer-review systems) make stronger judgments. However, in such a system, institutional cultures that avoid forceful professional judgments would be more likely to produce weak prior odds that are overridden by likelihood ratios (which data would drive). Such systems could satisfy educators' and teachers unions' concerns that personnel evaluation not rely entirely on test scores, because professional judgment would take precedence where it is exercised forcefully. But the ability of data to dominate weak prior judgments could also satisfy the concerns of policymakers dissatisfied with the unwillingness of administrators to make forceful judgments about ineffective teachers. A Bayesian algorithm provides a way out from the trust dilemma surrounding professional judgment and teacher evaluation: a "use it or lose it" approach to professional judgment is an alternative to either a dominant use of (imperfect) test data or a dominant use of (sometimes-reluctant) professional judgment, creating a possible operationalization of Shulman's (1988) *marriage of insufficiencies*.

Bayesian classifiers exist in practice, if not in education evaluation, and most of us have benefited by experiencing at least one—or rather not being aware of how we are benefiting. Most of the statistical research on the characteristics of Bayesian classifiers is in the field of machine learning, and many e-mail filters use Bayesian approaches to identifying spam (Graham, 2004; Sahami, Dumais, Heckerman, & Horvitz, 1998). Bayesian spam filters use the relative likelihood of several identifiable words or character strings from training sets to score a candidate e-mail as either likely spam or likely nonspam. As explained in some formal detail in the appendix, it is possible to use more than one data source in Bayesian classification (in spam filtering, multiple words). If one assumes that all data sources are independent of each other, then one calculates likelihood ratios and the classification comes from the product of a prior odds judgment (e.g., a professional evaluative judgment) with all of the likelihood ratios. Though an assumption of independence wreaks havoc with point estimates of most statistical inferences, there is considerable reason to believe that the yes/no decisions of a so-called naïve Bayesian classifier are not damaged much by an incorrect independence assumption (Domingos & Pazzani, 1997; Hand & Yu, 2001; Lewis, 1998; Rish, 2001; Zhang, 2001). Despite the research on the robustness of naïve Bayesian classifiers and the relative simplicity of calculation, they are not used in critical-decision frameworks where they could be of use, such as medical diagnoses. Recent research in medical diagnosis generally uses more complex Bayesian classifiers, which suggests that they can be substantially superior to more traditional diagnosis scoring methods (e.g., Biagioli, Scolletta, Cevenini, Barbini, Giomarelli, & Barbini, 2006).

Salient Data and Likelihood Functions

While Bayesian reasoning allows one to create a “use it or lose it” approach to evaluation in theory, that potential does not guarantee a practical “use it or lose it” algorithm. The utility of any such Bayesian system of evaluation depends on characteristics assumed in the prior section: the existence both of data and likelihood functions salient to the judgment of professional effectiveness. The limited curriculum coverage inherent in any test or assessment system is well-known (e.g., Dorn, 2007; Koretz, 2008), but to some extent, the use of likelihood functions in such a system would make the technical requirement of data use a little looser than the debates over test score use might lead one to believe. First, the explicit inclusion of professional judgment and the relationship between strength of judgment and influence over posterior odds (explained in the prior section) reduce the reliance on circumscribed sources of data. Second, likelihood functions could be constructed that accommodate measurement error; one category of candidate for such accommodation is the set of kernel likelihood functions (e.g., the likelihood of observing a datum plus or minus the measurement error, or a distributional likelihood of being observed at a certain percentile rank plus or minus a decile).[7]

It is thus the combination of data and likelihood function that needs to be salient to personnel evaluation, and the choice of a likelihood function for a specific source of data entails value judgments about our categorical judgments of effectiveness. Consider the type of judgment involved in the two choices about proficiency measures mentioned earlier—having an exact proportion of students judged proficient (or a kernel measure centered on an exact proportion) versus having that proportion of students or higher judged proficient. Identifying either an exact or a kernel measure as important implies that the precise place of student measures is more important to the judgment of effectiveness than a broad category such as "50% or more" and that distinguishing groups of students with 50% proficiency from all groups with 50% or greater proficiency is important in evaluation policy.[8] Either is a defensible position, but the consequences of the choices lead in different directions.

In addition to the choices of data and likelihood functions, one would need to identify an appropriate source of distribution for any likelihood function and a

[7] Domingos and Pazzani (1997) suggest dividing real-valued variables into five to ten bins, but the practical difference between a kernel and a quintile- or decile-based approach is beyond the scope of this paper.

[8] I assume here that *proficiency* is a meaningful construct. While the construct of proficiency depends on the validity of cut scores, which are always arbitrary (Glass, 1978), one could make the same argument as in the text with any ordinal measure chosen for the task at hand.

defensible categorization of data into the relevant bins. The assumption in the Bayesian reasoning presented above is that there already is a classification of teachers into different categories and an existing and known distribution of the data for each category. In reality, any chosen distribution is likely to depend on tentative proxy judgments: we tentatively divide a set of teachers into categories and use a sampled distribution of data on those teachers to create the likelihood functions. The political legitimacy of those proxy judgments and comparable samples would depend on the classification method. The tentative classification by administrators and teachers would be most acceptable in a political sense, but the consequences of such classification are also likely to make the task aversive for many who might be asked to participate.[9] While relatively simple in concept, the implementation of a Bayesian approach to evaluation involves both technical and political judgments.

Bayesian Reasoning and Additive Point Systems

The political benefit of a Bayesian approach is the possible construction of an evaluation system where professional judgment has a “use it or lose it” trait. That benefit is transferrable to an additive point system, with some restrictions on the structure of a point system. A point system with rigid weights or rigid maximums for the contribution of different components will not have the benefit of a Bayesian approach. A point system with more flexible relationships between different components can have

[9] Waving away the comparable-population question, one could predict a low response rate for teachers asked to judge the effectiveness of current peers in their schools, even if they are promised that their judgments would not affect personnel evaluation of their current peers *for that year*: their judgments would set the classifications used in later years and thus they would be responsible for setting the likelihood functions by which they and peers would be judged, at least in part.

the benefit of a Bayesian approach, and it is possible to construct a point system that is equivalent to a naïve Bayesian classifier. While the Bayesian equivalent point system is theoretical rather than a likely practice, the existence of an equivalent suggests what is necessary for a point system to capture the political benefit of a Bayesian approach.

Bayesian Equivalents in Additive Point Systems

As explained in the appendix, the conversion of a naïve Bayesian classifier to an additive point system requires a logarithmic transformation of the classifier's product of factors, a log transformation that creates a sum of log odds and log likelihood ratios. In theory, each contribution could be calculated based on the same likelihood ratios as described earlier, with a positive log likelihood ratio increasing the posterior odds of a target decision and a negative log likelihood ratio decreasing the posterior odds of a target decision. To preserve the equivalent reciprocal relationship among components, each component in the Bayesian equivalent point system must be unbounded on both the positive and negative ends. A strong prior statement in a Bayesian system is equivalent to a log odds further from 0 (either positive or negative) than any other component in a point system, and at least in theory, a point system can capture the political benefits of a Bayesian approach to evaluation.

Rigid Point Systems

In contrast to the point equivalent of a Bayesian approach, a point system with rigidly-bounded or –weighted components fails to capture the political benefits of a Bayesian approach to evaluation. In a point system with bounded components, there is no reciprocation among the components. The effective power of a set of judgments by professionals is entirely independent of the effective power of any other component. No matter how forceful or weak the judgment of an administrator or peer committee, the authority of all other components remain the same. Or, in practical terms, if the judgment of a principal is worth 50% of the potential points and most teachers receive identical scores, the influence of data remains the same as when a principal gives a range of scores to teachers. If evaluators responsible for one component of a point system are hesitant to make forceful judgments about weak teachers, other components do not become more important in compensation. If one believes in the abstract value of a particular component at the precise weight contained in a system, the advantage of such an approach is precisely the rigid authority of its components. However, such a rigid system depends heavily on the value judgments made in its construction and is unable to provide either type of assurance that would address the trust/mistrust dynamics in the debate over teacher evaluation.

Intermediate Options

However, a point-based system does not need to have fixed weights for component scales. With the removal of rigid bounds and weights from a point system, it is possible to capture the political benefit of a Bayesian approach to evaluation. If the weight for a component scale can vary, then one could introduce a reciprocal relationship between professional (supervisory and peer) judgments, on the one hand, and other sources of evaluative information such as student outcomes, on the other. The most important potential benefit of a Bayesian approach to evaluation is the political consequences of combining professional judgment and data in a way that gives more authority to professionals who are willing to make forceful judgments with the possibility of reciprocal authority for data when professionals are not willing to make forceful judgments. The practicality, advantages, and disadvantages of each approach described below will vary, and the purpose of describing some options is less to advocate for a particular approach than to illustrate a minimal range of approaches to point-based evaluation systems.

Weighting by component range. One such system would be a weighting of components by range, a literal “use it or lose it” formula. If a qualitative evaluation is worth up to half of the total points, but a set of evaluative ratings only spans half of the potential range, a “use it or lose it” policy could expand the weight of student-outcome data to fill the extra 25% of points not in the range of the qualitative evaluations. Such a system would be simple to explain and implement. It would also impose odd incentives to game the system, whereby a principal can insulate highly-rated teachers from the effect of test scores by giving extremely low ratings to a small number of teachers in a school.

Standardizing ratings by a central dispersion measure. A second approach would be an indirect reweighting by the transformation of both test-score data and professional evaluative ratings into standardized scores, in comparison to a central dispersion measure such as a standard deviation (within a relevant population). Suppose that the influence of professional evaluative ratings were set at twice the weight of ratings derived from test scores, after both are transformed into standard-deviation units. In a unit where administrators (or administrators and teachers, with peer review) provide a range of ratings to teachers, the extreme-valued ratings will be more influential than test scores, at both the high and low ends. On the other hand, if the professional evaluative ratings have no variation—where a standardized rating would be in the middle for all—then the test scores determine the end distribution. Administrators (and peers) could choose not to exercise their professional judgment in rating teachers, but such a choice would give compensatory authority to student outcome data.

Overdetermined total. A third approach could be a point system that theoretically overdetermines outcomes, with more than 100% of the potential range covered by the sum of components. For example, if professional judgment evaluation scores are worth 50 points, and data from student outcomes are worth 50 points, the range of the sum is 0 to 100. But if the range of sum scores is restricted to [20,80], each component's potential range spans 62.5% of the range for the total. The rationale for such a system would be that a system does not need to worry about extreme values that represent consistency between qualitative and quantitative sources of data: a teacher with high ratings in all categories is presumed to be highly performing, while a teacher with low ratings in all categories is presumed to be low-performing. It is in the middle of the range where the overdetermined sum has effect.

Conclusion

With the decision of the Obama administration to condition Race to the Top funds on the elimination of barriers to linking teacher and student test data, the weight of the U.S. political system is shifting towards linking teacher evaluations to test-outcome data. Some part of the policy discussion is focused on performance-pay policies and attendant choices, but the root importance of such a linkage is with regard to employment rather than pay: to what extent should teachers' jobs depend on test-score and other student outcome data?

This is not a new discussion, and the tensions involved in these policy debates will remain. Many teachers, administrators, and parents will oppose policies that place test scores in a dominant position, because they see tests as highly flawed and creating perverse incentives. School critics (including many parents) will oppose policies that result in uniform satisfactory evaluations for almost all teachers and see test-score use as an imperfect but justifiable tool to change evaluation practices. Without intervention, the likely outcome of these debates is a dichotomy of policies, with some school systems and states experimenting with crude uses of test-score data and other systems and states refusing to change, pointing to the inevitable problems with crude evaluation mechanisms.

This paper points in a different direction, using a Bayesian inference mechanism as a starting point. The description of unconventional algorithmic options is less to advocate for any of these approaches than to illustrate potential: one may be able to translate the most politically-valuable characteristic of a Bayesian approach to simpler algorithms, or at least algorithms that can be understood by a broad group of stakeholders. While one may not see formal Bayesian reasoning in personnel evaluation systems, there are some important lessons to take from the Bayesian approach and the parallel between the log transformation of conditional probability equations, on the one hand, and additive point systems, on the other. Most importantly, the political benefits from a Bayesian approach requires conscious construction in a point-based system of evaluation. While a Bayesian calculation of posterior odds explicitly creates a reciprocal relationship between prior odds and likelihood ratios, a point-based system with fixed weights/component contributions removes that reciprocal relationship. Adjusting weights, the use of standard-deviation-adjusted ratings, or overdetermined point systems are three methods to construct such reciprocal relationships, and it might be of significant benefit to explore such approaches.

While one probably could construct an evaluation system based entirely on a Bayesian approach, that is not necessary to gain the most important benefits: a range of technical solutions that provides reasonable incentives for all parties and a starting point for further development and local negotiations. Many teachers unions are unlikely to accept the use of test scores unless it is subservient to professional judgment, but other stakeholders are unlikely to accept the dominance of professional judgment without a backup method of evaluating teachers when the professional judgment is timid in judging weak teachers. A “use it or lose it” approach to professional judgment is a workable approach rooted in a Bayesian approach to inductive reasoning and with a few possible constructions within a point-based evaluation system.

Appendix

The relative-odds formulation of the Bayes theorem is the standard beginning point for naïve Bayesian classification, if one looks at relative odds of two possibilities A and B and considers them to be possible decisions. Let A be the need to intervene to help a poor teacher and B be nonintervention.[10] Then the relative odds of needing intervention versus nonintervention are

$$\frac{P(A|x)}{P(B|x)} = \frac{P(A)}{P(B)} \cdot \frac{P(x|A)}{P(x|B)} \qquad (2),$$

[10] The categories need not be exclusive: One could create additional categories such as recognition for merit, or dismissal, though the translation into an additive point system has difficulties with more than two categories.

where the first term on the right-hand side represents the prior relative odds of needing intervention and the second term is the relative likelihood of x given the two classifications under consideration (or likelihood ratio). One could reasonably interpret the first term as the judgment of relative need for intervention before gathering data $\{x\}$ and the second term as the relative likelihood of seeing the data under those relative judgments. If one had distributional information about $\{x\}$ for teachers needing and not needing supervisory intervention, where the data $\{x\}$ and the framing of the likelihood function were professionally salient (a question discussed below), then equation (2) allows one to adjust one's prior professional judgment by relevant data.[11]

In the cases of updating an administrator's initial professional judgment with data, the likelihood ratio is the key datum. In the case where the likelihood of data under condition A (intervention) is 6% and the likelihood of seeing the data under condition B (non intervention) is 1%, the likelihood ratio is $\frac{P(x|A)}{P(x|B)} = \frac{6\%}{1\%}$. The posterior (after-data-gathering) odds of needing intervention then become $\frac{3}{2} \cdot \frac{6\%}{1\%} = 9$, or a 90% posterior probability of needing remediation. But the data can also bump the prior judgment in the other direction. If 6% of teachers judged as needing remediation produce the data gathered but 8% of teachers judged as *not* needing remediation also produce the data gathered, the likelihood ratio is $\frac{P(x|A)}{P(x|B)} = \frac{6\%}{8\%}$. The posterior (after-data-gathering) odds of needing intervention then become $\frac{3}{2} \cdot \frac{6\%}{8\%} = 1.125$, or an only slightly greater than even odds (approximately 53% probability) of needing intervention.

[11] The question of appropriate data is discussed below.

Multiple Data Sources and Naïve Bayesian Classifiers

Consider first a Bayesian mechanism for adjusting professional judgment by two data sources rather than one, $\{x\}$ and $\{y\}$. Then the posterior odds become

$$\frac{P(A|x,y)}{P(B|x,y)} = \frac{P(A|x)}{P(B|x)} \cdot \frac{P(y|A,x)}{P(y|B,x)} = \frac{P(A)}{P(B)} \cdot \frac{P(x|A)}{P(x|B)} \cdot \frac{P(y|A,x)}{P(y|B,x)}, \tag{3}$$

which would require computational estimation of the likelihood ratios for interdependent data.[12] However, if $\{x\}$ and $\{y\}$ are independent, the last term becomes $\frac{P(y|A)}{P(y|B)}$ and

$$\frac{P(A|x,y)}{P(B|x,y)} = \frac{P(A)}{P(B)} \cdot \frac{P(x|A)}{P(x|B)} \cdot \frac{P(y|A)}{P(y|B)} \tag{4}$$

or, more generally, with $\{x_i\}$ for n independent variables,

$$\frac{P(A|x_1,x_2,\ldots x_n)}{P(B|x_1,x_2,\ldots x_n)} = \frac{P(A)}{P(B)} \cdot \prod_{i=1}^{n} \frac{P(x_i|A)}{P(x_i|B)} \qquad (5).$$

The concept here is that with a set of independent variables, or a series of data sources, one can repeatedly update the original professional judgment using the likelihood ratios of the different sources of data. The independence of the data sources is not an assumption likely to hold for most data sources in schools, but the simplified construct enables a direct comparison to point-based systems of evaluation, and there is some reason to believe that classifying algorithms are less vulnerable to inapt independence assumptions than real-valued estimators are (Domingos & Pazzani, 1997; Hand & Yu, 2001; Lewis, 1998; Rish, 2001; Zhang, 2001).

Log Transformation and Additive Points

A log transformation of equation (5) leads directly to a point-like system,

[12] Empirical Bayes estimation of the last term's quantities (likelihood of observing one variable given a prior and another variable) commonly involves Monte Carlo simulation using a Gibbs sampler. This paper is designed to provide a simpler introduction to the issues involved and assumes the identification of independent variables.

$$ln\frac{P(A|x_1,x_2,\ldots x_n)}{P(B|x_1,x_2,\ldots x_n)} = ln\frac{P(A)}{P(B)} + \sum_{i=1}^{n} ln\frac{P(x_i|A)}{P(x_i|B)} \qquad (6)$$

and if H= $ln\frac{P(A|x_1,x_2,\ldots x_n)}{P(B|x_1,x_2,\ldots x_n)}$, $\hat{h} = ln\frac{P(A)}{P(B)}$, and $h_i = ln\frac{P(x_i|A)}{P(x_i|B)}$, then equation (6) becomes

$$H = \hat{h} + \sum_{i=1}^{n} h_i \qquad (7),$$

which corresponds to an additive point-based system where a classification cutoff score for H corresponds to log relative posterior odds, $\hat{h}$ is the log odds of professional judgment for two categories, and each h_i is the log of a likelihood ratio estimated from $\{x_i\}$.

The transformation of a Bayesian updating system into a linear point-based system is not a statement that all point-based systems have an underlying Bayesian equivalent. The requirements here are steep: the correspondence of the first component to log prior odds (the qualitative professional judgment), the correspondence of additional components to a set of n independent sources $\{x_i\}$ (or a single data source $\{x\}$) and salient likelihood functions, and a cutoff score representing the relative odds at which a decision is appropriate. While it is possible to construct a point system in this manner (or to infer hypothetical, latent variables that operate in the way described here), the point of this exercise is not to suggest the construction of an explicit Bayesian-justified point system. Instead, the parallel can be a tool to explore the characteristics of any point-based system.

Weights. First, consider a set of weights $\{w_i\}$ for $\{h_i\}$, such that $H = \hat{h} + \sum_{i=1}^{n} w_i h_i$. Transforming this weighted linear formula back into equation (5), one can see that

$$\frac{P(A|x_1,x_2,\ldots x_n)}{P(B|x_1,x_2,\ldots x_n)} = \frac{P(A)}{P(B)} \cdot \prod_{i=1}^{n} \left(\frac{P(x_i|A)}{P(x_i|B)}\right)^{w_i}, \qquad (8)$$

where the weights $\{w_i\}$ become exponents for each data source's likelihood ratio. The consequence of exponential weights in equation (8) is partly dependent on the range of likelihood ratio for each source $\{x_i\}$ and also partly dependent on the threshold value of $\frac{P(A|x_1,x_2,\dots x_n)}{P(B|x_1,x_2,\dots x_n)}$ that would trigger a decision. If either the threshold value of the posterior odds is close to 1 or the likelihood ratio for a source x_i is close to 1, the exponential place of w_i becomes less consequential. The parallel in a weighted point system is similar: the weights for different components will not act in a linear fashion, but the effect of weights will depend on the implicit sensitivity of the threshold value for H and the range for each h_i. A threshold value for H that easily triggers a decision and restricted ranges for $\{h_i\}$ are associated with minimal effects of weights, while broad ranges for $\{h_i\}$ and classifications of H imply a highly nonlinear effect of weighting.[13]

Multidimensional scales. The consequences of multidimensional components follow from the nonlinear consequence of weighting. The concern about implicitly multidimensional scales for many measurement researchers is construct validity, but that may be less important in a type of evaluation that its designers intended to combine different types of sources. On the other hand, a multidimensional scale (or component of a point system) effectively sets weights for each dimension in ways that are not deliberate. A similar conclusion follows for sources of data that are not truly independent of each other. To the extent that a subset of variables $\{x_i\}$ underlying $\{h_i\}$ is not independent, collinear components of the subset of variables could be interpreted as a smaller set of variables that are differentially weighted. For example, if x_j and x_k are

[13] There is a similar effect of broad ranges in point-based grading systems. If the range for a course's component is on the order of the point range for a single grade, extreme values for a single assignment can shift a term grade by a letter grade, and that is equivalent to a nonlinear consequence of the assignment on the relative that a student will earn one grade as opposed to another.

linearly dependent so that $x_k = Kx_j$, one could replace x_j and x_k with $(K+1)\cdot x_j$, and $K+1$ operates as a weight with the consequences described above.

References

Azordegan, J., Byrnett, P., Campbell, K., Greenman, J., & Coulter, T. (2005). *Diversifying teacher compensation*. Denver, CO: Education Commission of the States. Retrieved July 26, 2009, from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED489329.

Baratz-Snowden, J. (2009, June). *Fixing tenure: A proposal for assuring teacher effectiveness and due process*. Washington, DC: Center for American Progress. Retrieved July 10, 2009, from http://www.americanprogress.org/issues/2009/06/teacher_tenure.html.

Behrstock, E., & Akerstrom, J. (2008, December). *Performance pay in Houston*. Rockville, MD: Center for Educator Compensation Reform. Retrieved July 26, 2009, from http://www.cecr.ed.gov/guides/summaries/HoustonCaseSummary.pdf.

Biagioli, B., Scolletta, S., Cevenini, G., Barbini, E., Giomarelli, P., & Barbini, P. (2006). A multivariate Bayesian model for assessing morbidity after coronary artery surgery. *Critical Care*, *10*(3), R94. doi: 10.1186/cc4951.

Bock, R., Wolfe, R., & Fisher, T. (1996). *A review and analysis of the Tennessee value added assessment system* [technical report]. Nashville, TN: Tennessee Office of Education Accountability.

Branigin, W. (2009, July 24). Obama launches “race” for $4 billion in education funds. *Washington Post*. Retrieved July 24, 2009, from http://www.washingtonpost.com/wp-dyn/content/article/2009/07/24/AR2009072402203.html.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103-130.

Fenton, N. E., & Neil, M. (2000). The jury observation fallacy and the use of bayesian networks to present probabilistic legal arguments. *Mathematics Today (Bulletin of the IMA), 36*(6), 180-187.

Glass, G. V (1978). Standards and criteria. *Journal of Educational Measurement, 15*(4), 237–261.

Graham, P. (2004). *Hackers and painters: Big ideas from the computer age.* Sebastopol, CA: O'Reilly Media, Inc. Graham's essay on spam filtering is also available at http://www.paulgraham.com/spam.html.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes: Not so stupid after all? *International Statistical Review, 69*(3), 385-398.

Howson, C., & Urbach, P. (2005). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago: Open Court Publishing.

Kaye, D. H. (1999). Clarifying the burden of persuasion: what Bayesian decision rules do and do not do. *International Journal of Evidence & Proof, 3*(1).

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Lewis, D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (pp. 4-15). Heidelberg, Denmark: Springer Verlag.

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics Journal, 43*, 147-163.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, *27*(3), 255-270.

Max, J. (2007, November). *The evolution of performance pay in Florida*. Rockville, MD: Center for Educator Compensation Reform. Retrieved July 26, 2009, from http://www.cecr.ed.gov/guides/summaries/FloridaCaseSummary.pdf.

Max, J., & Koppich, J. E. (2007, December). *Engaging stakeholders in teacher pay reform*. Rockville, MD: Center for Educator Compensation Reform. Retrieved July 26, 2009, from http://www.cecr.ed.gov/guides/EmergingIssuesReport1.pdf.

Potemski, A., & Rowland, C. (2009, April). *Pay reform in Minneapolis Public Schools: Multiple approaches to alternative compensation*. Rockville, MD: Center for Educator Compensation Reform. Retrieved July 26, 2009, from http://www.cecr.ed.gov/guides/summaries/MinneapolisCaseSummary.pdf.

Rish, I. (2001). *An empirical study of the naive Bayes classifier*. IBM Technical Report RC22230. Hawthorne, NY: T. J. Watson Research Center. Retrieved August 6, 2009, from http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). *A Bayesian approach to filtering junk e-mail*. AAAI Technical Report WS-98-05. Menlo Park, CA: Association for the Advancement of Artificial Intelligence. Retrieved August 6, 2009, from http://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-009.pdf.

Shulman, L. S. (1988). A union of insufficiencies: Strategies for teacher assessment in a period of educational reform. *Educational Leadership*, *46*(3), 36-41.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009, June). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved July 10, 2009, from http://www.widgeteffect.org/.

Wood, R. (1972). Review of *Bayesian Statistics* edited by D. L. Meyer and R. O. Collier, Jr. *The School Review, 80*(4), 629-640.

Zhang, H. (2001). *The optimality of naive Bayes*. Fredericton, NB: University of New Brunswick. Paper presented at annual meeting of the Florida Artificial Intelligence Research Society (Miami Beach). Retrieved August 6, 2009, from http://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf.