Education Policy Analysis Archives

Volume 10 Number 8 **January 28, 2002** **ISSN 1068-2341**

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Reactions to Bolon's "Significance of Test-based Ratings for Metropolitan Boston Schools"

Stephan Michelson
Longbranch Research Associates

Citation: Michelson, S. (2002, January 28). Reactions to Bolon's "Significance of test-based ratings for metropolitan Boston schools." *Education Policy Analysis Archives*, *10*(8). Retrieved [date] from http://epaa.asu.edu/epaa/v10n8/.

Abstract Several concerns are raised abut the procedures used and conclusions drawn in Craig Bolon's article "Significance of Test-based Ratings for Metropolitan Boston Schools" published in this journal as Number 42 of Volume 9.

Craig Bolon introduces his article about the Massachusetts Comprehensive Assessment System (MCAS), tenth grade mathematics tests, with a non-sequitur. "The state is treating scores and ratings as though they were precise educational measures of high significance," he tells us, but "statistically they are not." Nothing Bolon does leads to this conclusion. We do not know how "precise" these measures are, by any criterion. We do not know how "significant" they are, by any criterion. One wants the conclusions of an article to represent its contents. Bolon's do not. Whether scores measure something consistently is called "reliability," and Bolon's exposition implies highly reliable tests. [1]

Whether scores measure something else, something besides themselves, is called "validity."

Bolon has not said what concept success on these tests is supposed to imply, or that they don’t. He has not determined that these scores are invalid against some criterion. He finds them correlated with town income, a not surprising result that we have lived with for 35 years, since the publication of The Coleman Report. [2] Nothing in that correlation informs us about the validity of these test scores, their ability to predict some other attribute of the students who take them.

Bolon concludes that student characteristics, such as limited English proficiency, "all failed to associate substantial additional variance" over and above that accounted for by town income. As I will explain, that criterion is not helpful in determining if a variable should be in a model explaining test scores. And at any rate, it is incorrect. Bolon reaches that conclusion by dropping Boston schools from his data, weighting observations by size of school, and failing to take advantage of the very information he provides. He also claims that the data cannot display "performance trends," though he has not established that there are trends that the data fail to capture. The incentive systems the state will associate with school performance are not yet in place. Bolon cannot therefore say that schools have failed to react to them.

Finally, Bolon tells us that per pupil school expenditure *is* not well related to school performance. Or perhaps he is saying that is true over and above town income. It appears to be the lack of correlation between test scores and school expenditure that leads Bolon to conclude that schools cannot raise themselves when found to be at a low level.[3] In fact, per pupil expenditure is highly correlated with test score, and with income, also.

This comment is only about statistical analysis, about deriving information from data. I say "information" rather than "fact" because all statistical results are probabilistic, associations one can choose to accept or not on many grounds. On the surface that is what Bolon offers, statistically derived information. But his conclusions do not follow from his data. His procedures are suspect and his interpretations are incorrect. Surely the discussion requires better statistical analysis.

Data

Bolon has put together data from various sources, describing 47 schools in 32 towns around Boston. For this he deserves much credit, although one should not expect much from these data. Per pupil expenditure, for example, is presumably over an entire school system, and includes many items that no one would think to associate with a mathematics test score. Only if all students taking the test have always been in the school system in which we see them, at the tenth grade--and assuming that last year’s expenditures are representative of the last ten years of expenditures--would system-wide expenditure be a reasonable measure of resources expended on these students.

Within his data, Bolon recognizes that three Boston schools, entrance to which is predicated on an exam, "draw away many Boston students who tend to score well on achievement tests...." [4] But he neither includes that fact in his statistical study, nor does he investigate the effect of this "creaming" on other Boston schools. Let us do so quickly.

Boston schools, on average, do not score differently from non-Boston schools. Non-Boston schools do score higher (226.6 to 212.5 in 1999), but a difference this large or larger would be expected to occur by chance more than forty-four times in one hundred in a world in which the actual difference were zero. In Boston, the three exam schools out-score the ten non-exam schools (238 to 204.9) a difference that we would expect to occur by chance, assuming they were equal, fewer than two times in one hundred. [5] So a statistical examination of test scores should account for both those Boston schools that contain the cream (the variable I call "exam"), and those that contain only the remaining skimmed milk (the variable "bosnoex").

Bolon includes schools that "provide vocational education in the same facilities as academic programs." [6] He notes that "most students in vocational programs receive lower MCAS scores than students in academic programs" [7] when they are in separate schools, but nonetheless takes no account of vocational students when mixed with academic students. Bolon indicates this mix with an asterisk attached to twelve schools, none of them in Boston. I created the variable "voc," with a value 1 for schools that have a vocational program, 0 for schools that do not. A better variable would be the percent of students in each school who were "vocational." An even better variable would be the percentage of *test takers* in each school who could be categorized that way. If the indicator variable has a negative effect and is statistically unlikely from chance, when we assume it is structurally unrelated, then perhaps it is telling us something about these data that no other variable captures. It would not be *new* information, but how would one justify not taking into account old, correct information in asking what new information these data contain?

Replication

Furthermore, there is the problem of "specification error." Although this seems like a technical term, it is part of a larger inquiry about the language we use to describe our findings. The "specification" of a single equation regression is simply the variables, and their forms (linear, logarithmic, quadratic, etc.), including weights. Some variables might be important to specification even though their associated probability is high, because their inclusion changes the coefficients of other variables. One variable, though Bolon does not use it, does show some specification effect.

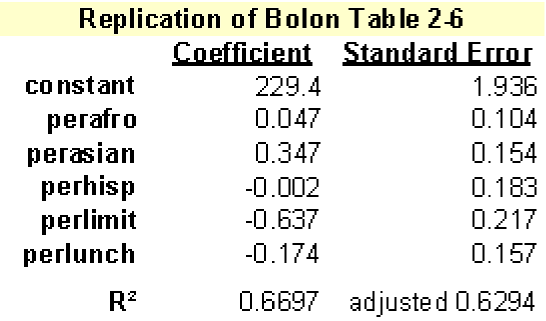
Bolon implies that school size is such a variable, but rather than including it, he weights observations by it. He is wrong on both counts: School size is not only not important for specification, weighting by it is incorrect. Weighting implies that there is more information in a large school than a small school. [8] If size affects score, we should determine that relationship directly (as a variable in the equation). Otherwise, Bolon’s observations are of schools, each one telling us as much as any other.

To this point it appears that Bolon has not taken advantage of his own data, failing to define three variables (exam, bosnoex and voc) that his own discussion implies should be important. If I am going to explain the consequences of these omissions, I should first replicate his results. Only if I do so can I assert that differences between us are due solely to the changes in specification that I will offer. It turns out that there is more amiss here than just the variables.

I placed my coefficients and standard errors into Excel, and let it round them to three places. Bolon uses what to me is awkward language for R^2, the most common statistic emanating from a regression equation. He says that a model "associates X percent of the statistical variance." Perhaps he is avoiding the more usual term, "explains." I am sympathetic to that desire, but "associates" is usually followed by the word "with." I find his phrasing unintelligible. If one were asked to estimate the average school test score of each school, knowing nothing about schools, the best strategy is to estimate the mean of all schools every time. R^2 essentially tells you how much better than that you can do (how much closer to real values your estimates are) if you generate an estimated score from the regression equation. R^2 can vary from zero to 1.00.

In an exercise in which Bolon will use different numbers of explanatory variables, he should report the adjusted R^2, which makes the statistic pay a penalty for using more variables. The easy strategy to increase R^2 is to use all the variables you have. R^2 itself can never decrease when variables are added, an important item to remember as Bolon changes data on us. Adjusted R^2 can decrease, telling you that you have paid too high a price for whatever information your last variable has added. It can even become negative. I will return, below, to how Bolon appears to be using R^2 to create his specification. In Table 1, a replication of his Table 2-6, to the three places Bolon reports, we agree exactly.

Table 1



Replication of Bolon Table 2-6

	Coefficient	Standard Error
constant	229.4	1.936
perafro	0.047	0.104
perasian	0.347	0.154
perhisp	-0.002	0.183
perlimit	-0.637	0.217
perlunch	-0.174	0.157
R^2	0.6697	adjusted 0.6294

The prefix "per" represents "percent." I believe we are looking at the percent black, Asian and Hispanic in the school, not necessarily taking the exam. Similarly, "perlimit" is percent of the school with limited English fluency (whether defined consistently from school to school I cannot say), and "perlunch" is percent receiving free or reduced cost lunch. A variable we will encounter below, "percapy," is not a percentage. It is per capita income in 1989 (1990 Census data). Short of administering a questionnaire to each student--and then dealing with the accuracy of the responses--these are the kinds of variables we have in education research, somewhat distant from what we would like to measure.[9]

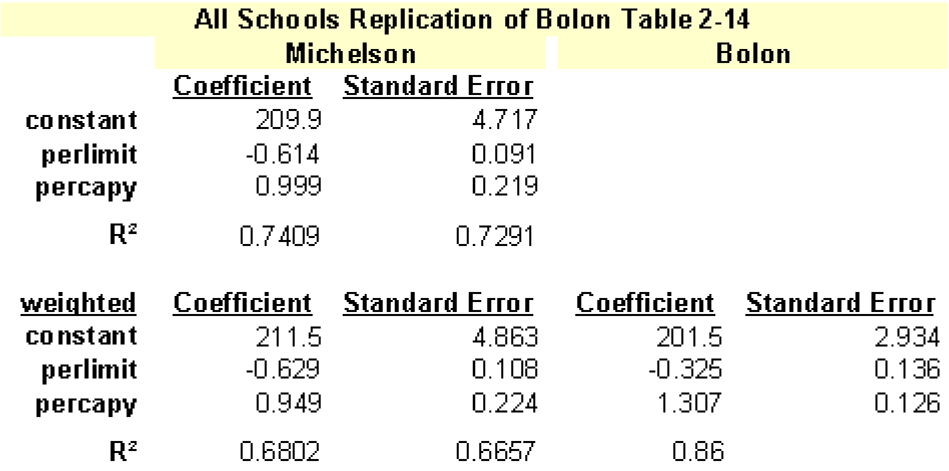
I will present R^2 and adjusted R^2 from left to right in the following tables, without including the word "adjusted." I could show a similar table (replicating Bolon's) for his Table 2-7. In Table 2-9, Bolon and I disagree in the last decimal place shown of four coefficients. It is possible that Bolon has data to more decimal places than he has reproduced. [10] This possibility is suggested by coefficients for per capita income in his equation in Table 2-10. Mine is 1.108, his is 1.104. Not consequential, but not a rounding difference, either.

A Change of Method

I think we can agree that my results are what Bolon would derive from the data he has provided. Just after his Table 2-13, Bolon makes a turn that he does not justify, and that I see no reason for. He tells us in Table 2-13 that the unadjusted R^2 for the three factor equation in Table 2-10, in 1999, is .80. That is, the residual variance of the difference between school scores and his estimate thereof is 20 percent of the value it was when he estimated every school to have the mean score of all schools. He has, in more traditional language, "explained" 80 percent of the variance in scores. Yet he will tell us, in Table 2-15, that a two factor model produces an R^2 of

.86, and that he needs only per capita income to generate an R^2 of .84. Higher unadjusted R^2 from fewer variables? Something else has changed.

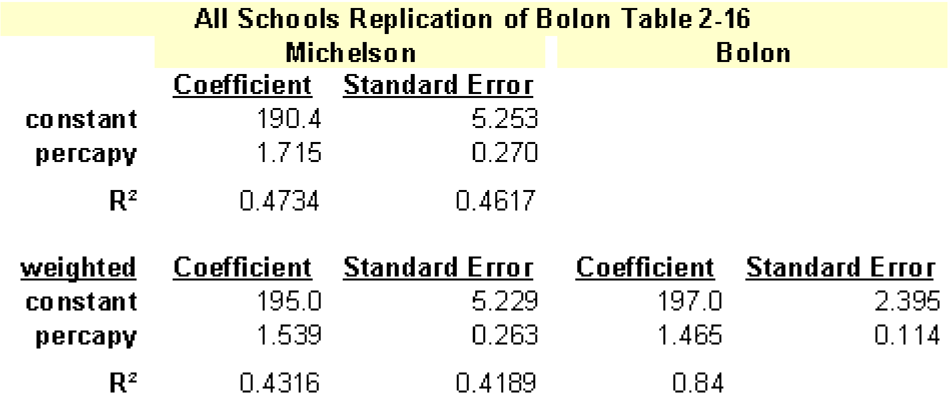
Table 2



All Schools Replication of Bolon Table 2-14				
	Michelson		Bolon	
	Coefficient	Standard Error		
constant	209.9	4.717		
perlimit	-0.614	0.091		
percapy	0.999	0.219		
R²	0.7409	0.7291		
weighted	Coefficient	Standard Error	Coefficient	Standard Error
constant	211.5	4.863	201.5	2.934
perlimit	-0.629	0.108	-0.325	0.136
percapy	0.949	0.224	1.307	0.126
R²	0.6802	0.6657	0.86	

One difference lies in the statement just before Table 2- 14, that from now on regressions are calculated "with schools weighted by numbers of test participants." Bolon does not tell us why he has changed his specification that way. If this is how regressions should be run on these data, why did he not start out doing so? Nor does this change in procedure alone allow me to replicate his results.

Table 3



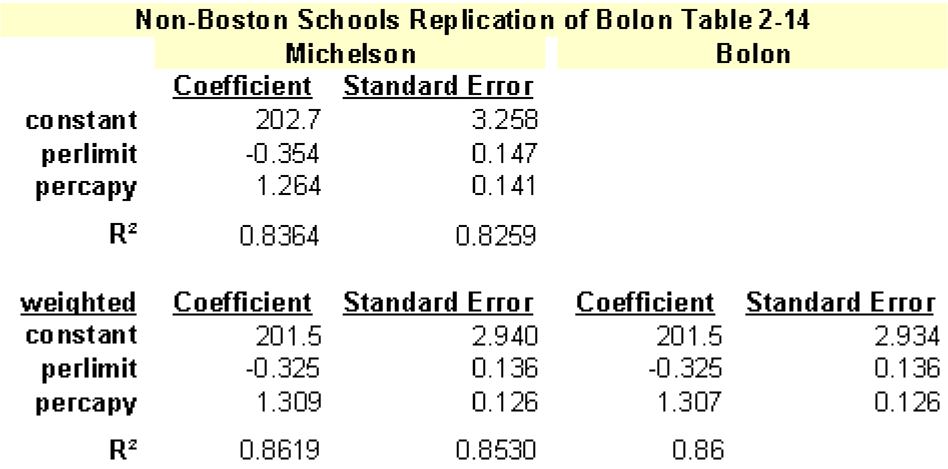
All Schools Replication of Bolon Table 2-16				
	Michelson		Bolon	
	Coefficient	Standard Error		
constant	190.4	5.253		
percapy	1.715	0.270		
R²	0.4734	0.4617		
weighted	Coefficient	Standard Error	Coefficient	Standard Error
constant	195.0	5.229	197.0	2.395
percapy	1.539	0.263	1.465	0.114
R²	0.4316	0.4189	0.84	

In Tables 2 and 3 I have weighted by the adjusted number of test takers, as listed in Bolon’s Appendix Table A3-1. Weighting makes little difference in the coefficients or the R^2. Weighting surely should be justified, which I believe it cannot be in this context, but neither is it of any consequence.

As implied by my table titles, and R^2 far below the values Bolon reports, what Bolon has done is drop some schools. Boston schools. He does so, partly, "because the students who score well on school-based standard tests are selected for admission to the three exam schools," and for other reasons mostly confined to a footnote. [11] As we know, the exam schools can easily be represented by a variable. He notes that attributing the same per capita income to all ten schools in Boston does not allow that variable to explain differences in test scores among those schools. True, but deleting variation because you cannot explain it is guaranteed to both raise R^2 and leave you ignorant. For Bolon to conclude, later, that per capita income is the only factor needed to explain most variation in test scores, is to ignore that he has deleted 21 percent of his data to achieve that result.[12] It is not true of his initial data set. He does not feel compelled to make the same adjustment in Lynn, Newton or Quincy, all of which have the same per capita income applied to more than one school. Nor is it necessary to adjust the data at all.

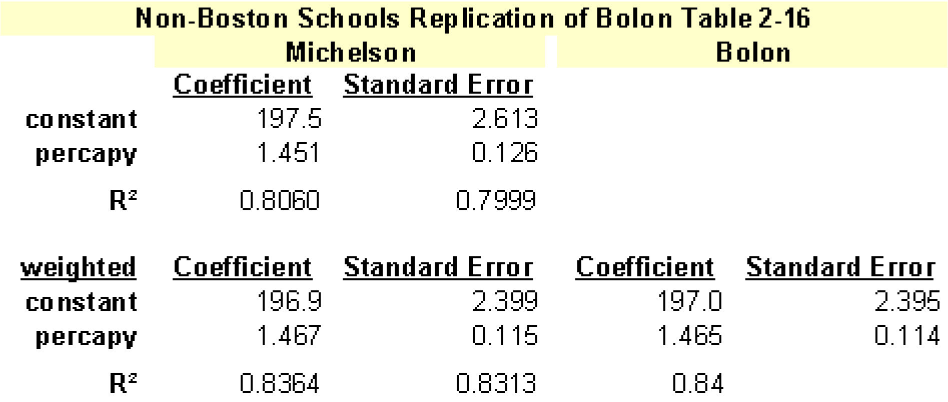
One adjustment Bolon might have made is to have one observation per town, by creating a weighted average of scores (as he tells us how many test takers there are per school). In fact, I think there is no school effect to be found in these data--though there are pupil effects--and therefore a per town analysis would have been best. I will return to that view. Another answer is to use the variables I suggested above, which assume that exam schools will have higher than average scores because they have selected students on that basis, and that Boston's non-exam schools will have lower than average scores because the students they would ordinarily enroll have been snatched from them by the exam schools. First, I will show that dropping the Boston schools, not weighting, is the key to Bolon's analysis:

Table 4



Non-Boston Schools Replication of Bolon Table 2-14				
	Michelson		Bolon	
	Coefficient	Standard Error		
constant	202.7	3.258		
perlimit	-0.354	0.147		
percapy	1.264	0.141		
R^2	0.8364	0.8259		
weighted	Coefficient	Standard Error	Coefficient	Standard Error
constant	201.5	2.940	201.5	2.934
perlimit	-0.325	0.136	-0.325	0.136
percapy	1.309	0.126	1.307	0.126
R^2	0.8619	0.8530	0.86	

Table 5

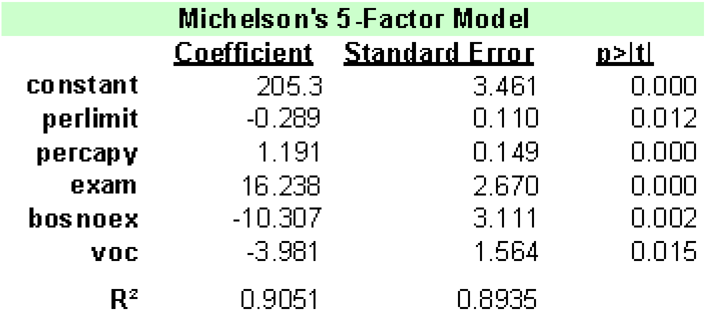


Non-Boston Schools Replication of Bolon Table 2-16				
	Michelson		Bolon	
	Coefficient	Standard Error		
constant	197.5	2.613		
percapy	1.451	0.126		
R^2	0.8060	0.7999		
weighted	Coefficient	Standard Error	Coefficient	Standard Error
constant	196.9	2.399	197.0	2.395
percapy	1.467	0.115	1.465	0.114
R^2	0.8364	0.8313	0.84	

As earlier, here in Tables 4 and 5, comparing equations vertically on the left, there is little difference between weighted and non-weighted results, again raising the question why Bolon bothered to change his method at this point. Reading the weighted equations left to right on the bottom, there are some differences between us in the last decimal place that cannot be attributed to output rounding. Essentially I replicate Bolon, which means that he has decided that Boston schools do not contribute to our knowledge about test scores in Boston-area schools.

He is quite wrong about that. Consider this five factor "model," which I present unweighted:

Table 6



Michelson's 5-Factor Model	Coefficient	Standard Error	p>|t|
constant	205.3	3.461	0.000
perlimit	-0.289	0.110	0.012
percapy	1.191	0.149	0.000
exam	16.238	2.670	0.000
bosnoex	-10.307	3.111	0.002
voc	-3.981	1.564	0.015
R^2	0.9051	0.8935	

I have added a column. I do not find the standard error a very informative statistic. Without the t-distribution at hand, one cannot translate the coefficient and standard error to probability. Many people report the t statistic, which is the ratio of the coefficient to its standard error, but it suffers from the same problem. I prefer to report the probability itself, as that is what everyone receiving other information is trying to derive.

Specification

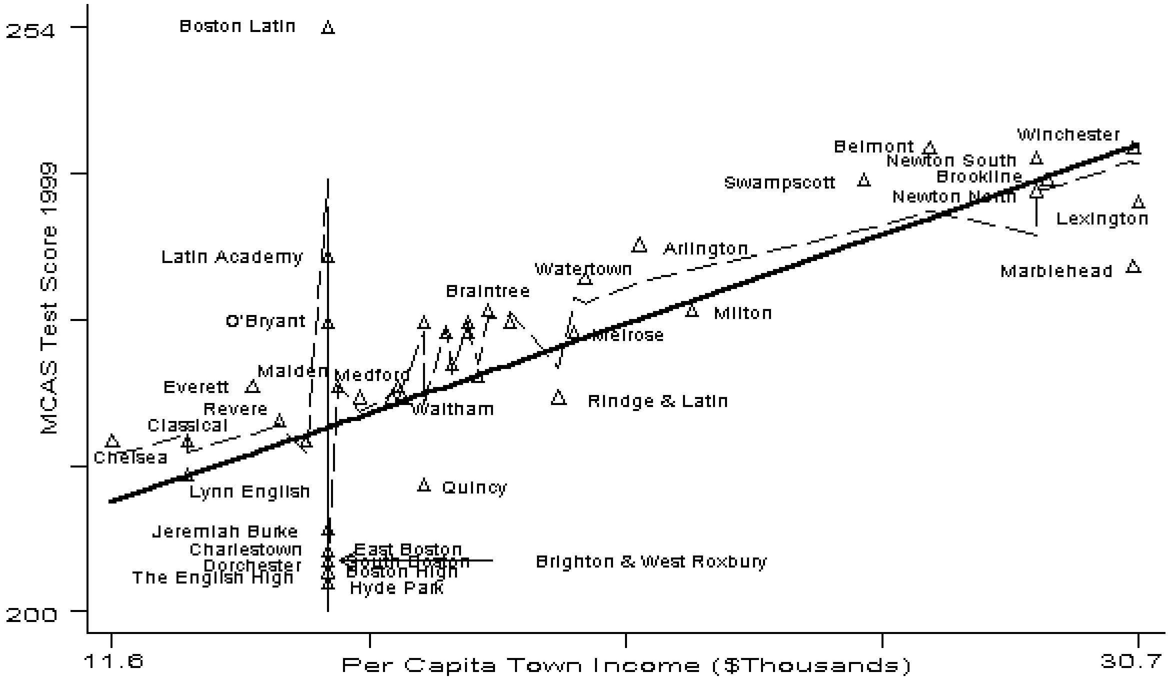
Bolon is much too attached to reporting "significance, and eliminating variables "not significant at $p < .05$." That is what I call a "mechanical" approach to statistical inference, one a machine can do as well as a human. Probability is not like an on-off light switch. It is more like a dimmer. There is more or less of it. $p = .06$ should not be set aside as "not significant." It requires interpretation. I would accept a probability for the vocational variable higher than for other variables, because it masks the variation (the percent of exam takers who are vocational) that it should have. The equation in Table 6 does not need this explanation. [13] It would have been a superior model with higher probabilities, but its probabilities meet Bolon's apparent criterion.

The three classification variables included here, derived from Bolon's data but not presented as such by him, are extremely powerful explainers of variation in school test scores. My equation achieves a higher R^2, even adjusted, than any unadjusted R^2 Bolon reports. I don't say that should be the single criterion for choosing a model, but it does seem to be one of Bolon's, and is a good starting point. Retaining all the data and achieving a higher R^2 than Bolon's model did, after his dropping data for the sole purpose of increasing R^2, surely argues that my model is more informative than his.

If statistical comparisons are too ornery--I often think they are--a picture should make my point. Let's compare the ability of Bolon's "one-factor model" and my "five factor model" to predict the data, to estimate the actual data points from applying the equation to input data.

In Figure 1, test scores from 1999 are plotted for all 47 schools, against per capita income. Triangles represent data points. Estimated points are implied by the lines that join them. The solid line is the one Bolon would have us believe best describes the data, a regression on per capita income alone. The dashed line follows estimates made from my five factor equation, Table 6. Does Bolon's one factor model capture the essence of this data set? I don't think so.

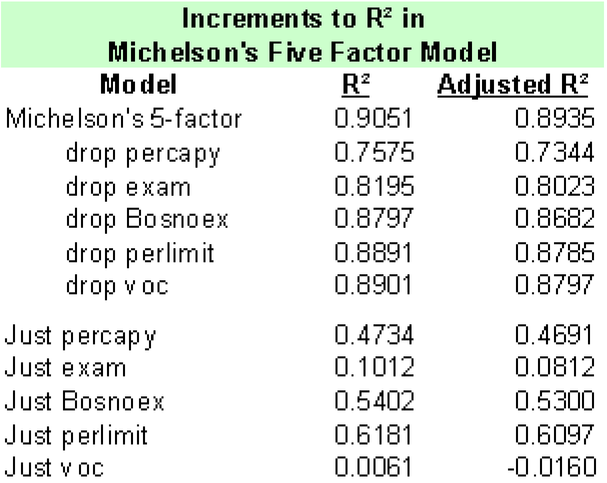
Figure 1



Boston schools lie on top of each other, because per capita income is defined only by town. Bolon's regression line, of course, cannot distinguish among them. However, my regression distinguishes all except Boston Latin very well. It is true that the non-exam Boston schools cannot vary in per capita income, but neither do they vary much in test score. Although not always in the same place, the ten Boston non-exam schools had the ten lowest school averages in all three years of Bolon's data. Boston Latin always had the highest score. Bolon's equation—or, to be more precise, my equation using Bolon's specification on all schools—under-estimates scores of low income towns and over-estimates scores of high income towns. In short, he presents not only an inaccurate but a biased view of the net relationship between per capita income and test score.

Besides looking to coefficient probability and R^2, how does Bolon (apparently) select his model? In Table 2-12 he presents R^2 for each combination of three factors. In Table 2-15 he asks how much increase in R^2 he can obtain from adding alternative other variables to his two factor model (now based on only 34 schools, weighted). He tells us that one factor, per capita income, gives him an R^2 of .84, so why bother with limited English proficiency when it increase R^2 to only .86? This is a political decision. Bolon wants to say that a model with no variable describing school or child characteristics tells us as much as we can know about these data. He wants to argue that school test data that relate only to town income cannot be valid, though this is not how validity is measured. There is no statistical justification for dropping other variables.

Table 7



Increments to R^2 in Michelson's Five Factor Model

Model	R^2	Adjusted R^2
Michelson's 5-factor	0.9051	0.8935
drop percapy	0.7575	0.7344
drop exam	0.8195	0.8023
drop Bosnoex	0.8797	0.8682
drop perlimit	0.8891	0.8785
drop voc	0.8901	0.8797
Just percapy	0.4734	0.4691
Just exam	0.1012	0.0812
Just Bosnoex	0.5402	0.5300
Just perlimit	0.6181	0.6097
Just voc	0.0061	-0.0160

In Table 7, I show, first, how much the addition of each variable in my Five Factor Model adds to R^2. The number after "drop" is the R^2 from the remaining four variables. Thus, for example, per capita income adds .1476 to R^2, because R^2 is .7575 without it, and .9051 with it. Per capita income alone generates an R^2 of .4734, from the bottom half of Table 7. In other words, .3258 of the variation "explained" by per capita income is also explained by other variables. That is, 69 percent of the apparent explanatory power of per capita income may or may not be due to per capita income. It is "shared" with other variables. One might say that the variable "exam" explains little--an R^2 = .10 when it is the only variable in the equation--but only 15 percent of that is shared with other explanatory factors. Limited English proficiency explains more than any other single variable, but shares 97 percent of that variation with other variables. Is this a reason to drop this variable?

Not statistically. When Bolon drops perlimit, he allocates its shared explanatory variation to the remaining variables, in his case to percapy, by fiat. In the regression context, we just do not know whether it is per capita income or English proficiency, or both, or neither (they may both be proxies for something else) that is associated with variation in test scores. We let the multiple regression statistics tell us whether adding perlimit is worth the price of using up a degree of freedom. The variable describing limited English proficiency belongs in the model, even though 97 percent of the variation in test score that it explains can be explained by other variables. We can believe, with about as much confidence as one ever can get from data such as these, that where there are more students with limited English proficiency, the school's average test score will be lower.

The only exception to this arithmetic, where adding the R^2 of the equation with just one variable to the R^2 of the equation with all other variables produces a larger R^2 than we know all five produce, is vocational education. Adding its R^2 alone to the R^2 without it falls short of the R^2 we know we get from five factors. Vocational education displays a specification effect. Its presence in the equation increases the contribution to R^2 made by other variables. Bolon should have defined and utilized a "voc" variable regardless of its contribution to R^2.

Interpretation

We might infer that it is the students with limited English proficiency themselves who are causing a decline in the average, through their low scores. That is an inference Bolon is careful not to make. There is such a thing as the "ecological fallacy," or the "Simpson paradox," in which group data lead to incorrect inferences about individuals. But fear of that mistake should not deter someone from making an informed judgment. Studies in the 1960s and 1970s showed that states with more blacks had lower income, and states with higher education levels had higher income. It was not wrong to infer that blacks earned less than whites, and that higher educated people earned more than lower educated people, even though it *was* wrong to infer (at that time, though many did) that if black education levels rose they would earn considerably higher incomes. I think Bolon's data is quite sufficient to ask if the MCAS test is fair to persons with limited ability in English (and, presumably, Spanish, as tests are available in that language).

The same is true with vocational students. There could be many reasons why schools with vocational programs score lower than strictly academic schools, without the vocational students being the direct cause. But, armed with his other study of strictly vocational schools, Bolon surely can infer that it is the vocational students themselves who are scoring lower. This raises the question: What purpose is served by vocational students taking this academic test? Bolon's study does *no*t answer this question, because it is not a validity study. But this is an example of the kind of useful inquiry Bolon could have provided.

We can conclude that two student descriptive variables appear important, an interesting if expected finding from aggregate data of this sort: limited English language ability and a vocational curriculum. In addition, we can conclude that creaming does matter, in the sense that concentrating all the best students in three schools concentrates the worst students in the remaining schools available to them. Ultimately, we can interpret these data as telling us nothing about schools, but much about students.

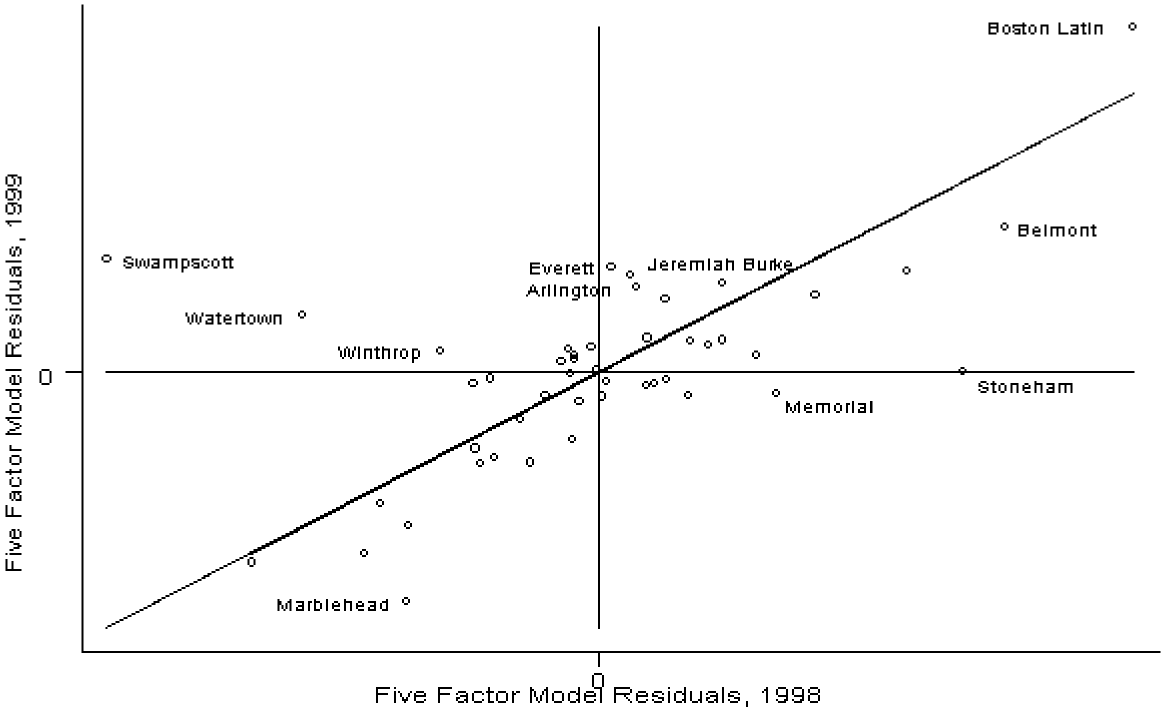
It was conventional wisdom in 1955, when I graduated from Brookline High, that aside from Boston Latin, one could find the best greater Boston area students in Brookline and Newton, and maybe Lexington, largely because their parents had moved to these towns for that reason. Ten years later research declared what we knew to be true, though it was "controversial" in professional ranks: high scoring schools were so because they had high scoring students. And high scoring students had two characteristics: By and large they came from wealthier households, and they associated with each other in wealthy towns. Bolon's finding to this effect is surely correct, and welcome, but not new information, and hardly impugning the MCAS test on which it is based.

Residuals

Bolon estimates his one-factor model in three years, and concludes "that schools scoring higher than predicted tend to increase scores in successive years, and schools scoring lower than predicted tend to decrease scores." [14] This is not a correct interpretation from his own data and results. Residuals do not show the highest scoring schools scoring even higher from year to year, or the lowest scoring schools scoring even lower. The correlation of residuals means that they are in the same place every year. Furthermore, 1998 scores do not predict 1999 scores for the lowest scoring schools. The lower scoring schools have randomly higher or lower scores the next year. Although 1998 scores appear to predict higher 1999 scores for the highest scoring schools, that result is due to Boston Latin only. Except for Boston Latin, the higher scoring schools in 1998 have somewhat lower scores in 1999.

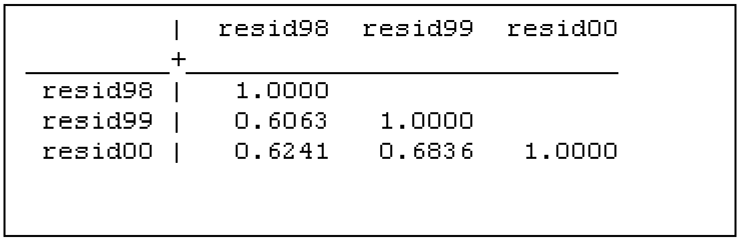
Figure 2 is a plot of residuals in 1999 against residuals in 1998, from my 5-factor model:

Figure 2



As I did also in Figure 1, I have named some schools, the principle being to inform but avoid clutter. In general, schools line up the same way every year. The residuals from 1998, for the top fifteen schools, predict the residuals from those schools in 1999 with a very small constant and a coefficient of .995. The residuals from the bottom fifteen schools in 1998 predict their 1999 residuals with a zero constant and a coefficient of 1.055. That is, we can predict the residual for the top and bottom schools almost precisely from the prior year data. The same does not hold true in the middle. Overall the residuals are correlated, as we would expect them to be:

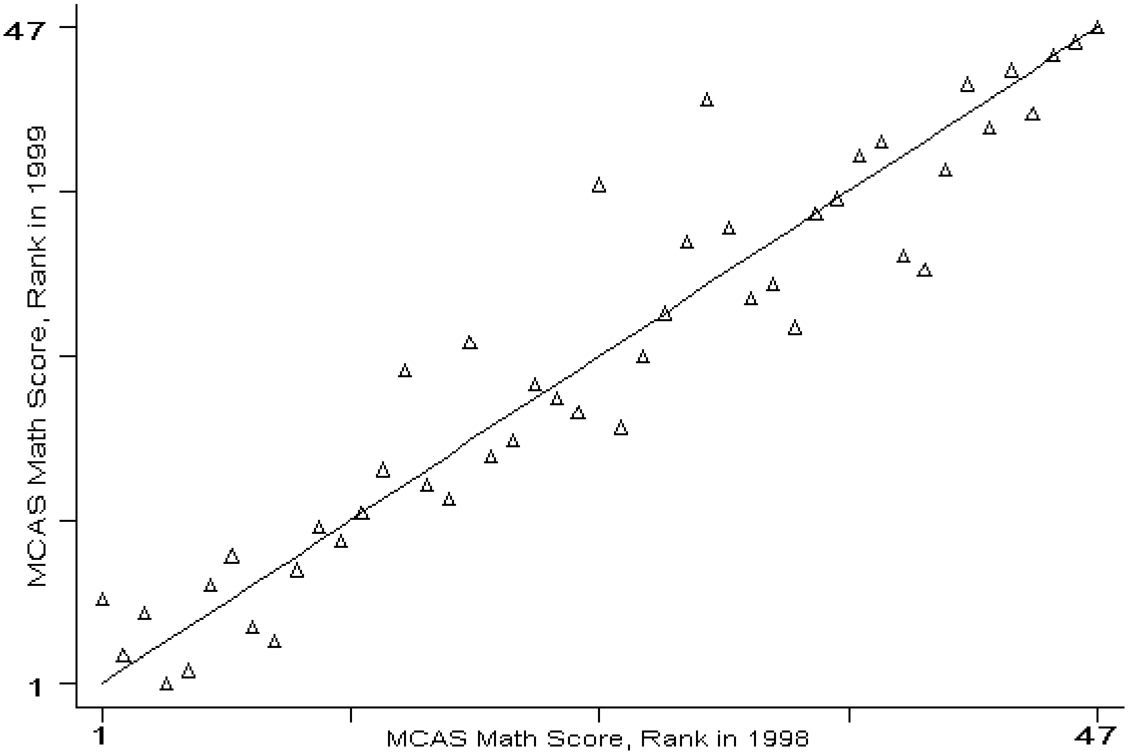
Table 8



```
          |  resid98   resid99   resid00
----------+------------------------------
  resid98 |   1.0000
  resid99 |   0.6063    1.0000
  resid00 |   0.6241    0.6836    1.0000
```

More importantly, raw scores are also correlated across time, as indicated above in noting that the Boston non-exam schools always score at the bottom. Figure 3 is a picture of the ranks of the scores, from lowest to highest, in 1999 and 1998:

Figure 3



The best schools--which is to say the best students--are consistently best, and the worst are consistently worst, and in the middle there is some moving around, but the middle schools seldom appear best or worst. This is startlingly true, and largely unrecognized, of most reliability measures. A test with a high reliability can be very poor in the middle, where it is used to hire or not, to pass or to fail. It is likely that this same variability applies to individual students, and that reliability figures offered in support of the MCAS test do not describe the unreliability at exactly the pass-fail scores where it is most important to be correct.

The Town View

As noted above, another solution to Bolon's dilemma, that some variables described towns, not schools, would have been to aggregate all data by town. The assumption that would allow one to do this is that there is no school effect, for example that Boston Latin students score highest because they were pre-selected, not because they were better educated. Nothing in these data argues against that view.

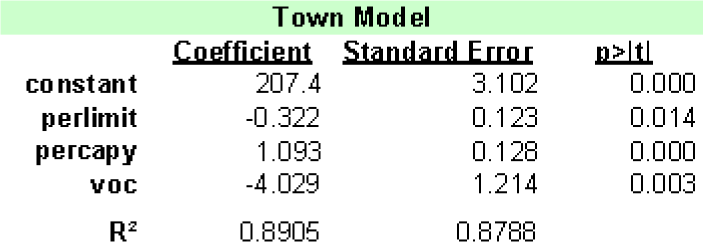
Bolon chose a different route:[15]

> My summary analysis is based on a one-factor model for metropolitan Boston communities that each operate only a single academic high school, weighted by numbers of students tested. The effects study showed that "Per-capita community income (1989)" was the dominant factor in predicting 1999 school-averaged tenth-grade MCAS mathematics test scores. All other factors made only small contributions to predictions with much lower significance.

That is, Bolon's final comments are based on 60 percent of his initial data, 28 observations. It is hard to take Bolon's conclusions as seriously generalizing the results of a study of 47 schools from 32 towns. I will suggest below that only one town--not one selected by Bolon--might have been dropped to increase our understanding of existing relationships.

One of the interesting results one can obtain from analyzing towns is that property value (measured per capita in Bolon’s data) appears to be a function of both income and educational expenditure, but not MCAS test score. This finding sort of follows conventional wisdom, that people pay a premium for property where there is a higher focus on education, though usually we expect a more direct relationship with test scores.

Table 9



Town Model			
	Coefficient	Standard Error	p>|t|
constant	207.4	3.102	0.000
perlimit	-0.322	0.123	0.014
percapy	1.093	0.128	0.000
voc	-4.029	1.214	0.003
R^2	0.8905	0.8788	

Following Bolon’s main concern, the best model from town data is quite familiar. I show its coefficients in Table 9. It is my five factor model less "exam" and "bosnoex," which have been aggregated into a single Boston observation.

I contend that the highest scoring students are exactly who we would expect them to be, those from highest income places, modified by those student characteristics that we expect to generate lower scores, such as vocational curriculum and limited ability in English. Boston fits right in. The most "influential" observation, by far, is Marblehead. Without that observation, the adjusted R^2 = .9221, the coefficient for perlimit is -2.73 (p = .011), and the coefficient for percapy is 1.247 (p = .000). The coefficient for voc is close to that in Figure 9. Thus the general effect of limited English ability (as well as income) on MCAS score is quite a bit larger than Bolon suggests, or that my model on all towns suggests. If we are going to delete data, let us do so to improve the generalizability of the results. When we do that, we emphasize the importance of the points I have made (that "voc" and "perlimit" need to be in any model explaining test score in these data).

Final Remarks

There is nothing wrong with analyzing a set of data and concluding "Nothing new here, looks like these relationships usually do." Unfortunately, such findings are seldom published. So we systematically lose confirming information, or broadening information (that relationships found elsewhere apply here) that Bolon could have provided. There *is* something wrong with saying one has analyzed data and found that "community income swamped the influence of the other social and school factors examined" when that conclusion is drawn from failing to utilize the information at hand, and dropping two-fifths of the observations. Bolon concludes: "Large uncertainties in residuals of school-averaged scores, after subtracting predictions based on community income, tend to make the scores ineffective for rating performance of schools." What are these "uncertainties?" Why should the state care about the residual from Bolon’s model? That his model produces residuals that do not correlate perfectly from year to year implies nothing about the effectiveness of rating schools that, as shown here, do pretty much line up the same every time we look.

He goes on: "Large uncertainties in year-to-year score changes tend to make the score changes ineffective for measuring performance trends." Does he mean "large *variation*?" Variation may not be uncertainty. His attempt to define "large" from published student reliability data is innovative and bold but, though I have not dwelled on it here, not convincing in a study of school averages. Wouldn’t "trends" be measured by consistency of direction? Unless Bolon can tell us that there are trends which these data fail to pick up, what he is saying is that "trend" studies should account for random variation, that to establish a "trend" means to exceed variation from changes in the students and tests and random factors from year to year. Well yes, of course. Is he saying no school could meet this criterion? Quite the contrary, he says that too many do. Where in all other respects, when Bolon finds a result inexplicable by chance, he accepts it as "significant," in his "trend" study he rejects the results just because they appear to be inexplicable by chance.

When he concludes that "tenth-grade MCAS mathematics tests mainly appear to provide a complex and expensive way to estimate community income," Bolon is ignoring the fact that MCAS tests are designed to measure individual achievement. If they do that job, then Bolon will find that individuals who do well traditionally are found with others who do well. His finding of a correlation of score with income does not argue that the test fails to measure understanding of mathematics. Does Bolon not believe that more of the "best" academic students, in general, attend Boston Latin, or Brookline, Newton, Lexington and a few other elite schools, than other schools? Does he not believe that the three Boston exam schools "cream" the best students from other schools, leaving these other schools the lowest scoring in the area?

What is *wrong* with finding that high scores in attributes that produce wealth are correlated with the wealth they produce? This is one reason my parents moved to Brookline, and surely one reason Bolon lives there, too. Why does Bolon not want to emphasize the more important findings that students who have difficulty with English, or with an academic environment, do not do well on this academic test? I do not know why the Massachusetts Department of Education wants to impose this test on all students, or why it wants to deny a high school graduation to those who fail it. However, nothing in Bolon’s article argues against them. Bolon is free to disagree with the state (as would I), but saying that there is a high correlation between test score and community income, not news in the twenty-first century, does not support his position.

Notes

[1] Bolon would disagree. He questions their reliability in Tables 2-1 and 2-2 based on year to year changes in school averages. I discuss reliability to some extent below. I agree that Bolon raises good questions, but I disagree that Bolon can criticize test reliability, as understood by test makers, in this manner.

[2] Coleman, James S., E.Q. Campbell, C.J. Hopson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York, Equality of Educational Opportunity (Washington, DC: U.S. Department of Health, Education and Welfare, 1966).

[3] I see no rationale for utilizing land value as an independent variable predicting test scores. By the middle 1970s, urban planners had determined that property values reflect school scores. People want higher scoring friends for their children. No one has shown the research that says people act this way, or their behavior, to be wrong. Bolon has the direction of causality reversed even as a hypothesis.

[4] Bolon, last paragraph of Section 1, Part B.

[5] The probabilities are from the t- distribution applied to the difference in scores divided by the joint standard deviation (the square root of the sum of the variances). However the Boston-other comparison is two tail, because we have no hypothesis about which schools would score higher, whereas the exam-not exam comparison is one-tail, where a prior hypothesis about the direction of the difference is clear.

[6] Third paragraph of Section 1, Part B. He also says "I choose to include such schools in these studies while noting their special character." In no part of his analysis is their "special character" noted, although it could have been.

[7] First quote from Bolon p. 4, second from p. 6.

[8] I was plaintiffs' statistical expert in the "Comparable Worth" litigation in the state of Washington in 1983. I weighted the salary structure of jobs by the number of employees per job, because salaries in larger jobs were in fact set more carefully than those in smaller jobs. The larger jobs contained more information about whether gender appeared to be a factor in determining salary, so weighting was appropriate.

[9] As Bolon says, "the data generally available for test score research fail to capture much of the critical information needed to understand development of cognitive abilities and educational achievement in the settings of public schools." Section 1, Part D, first paragraph.

[10] Craig Bolon kindly sent his data to me. I did not examine each item, but I did assure that the correlation between his data and mine, for each variable, was 1.0000. It is conceivable, but unsettling if true, that his program Statistica, produces different results from mine, Stata.

[11] Bolon, second paragraph after Table 2-13.

[12] Bolon equates R^2 with "accuracy." "In an attempt to improve accuracy of the model in Table 2-14, schools with residuals from the two-factor model for 1999 that were greater than two standard deviations were dropped." (at Table 2-18). Influence, not residual size, can be used to delete variables to increase the generalizability of results. See below. All that is accomplished by deleting large residuals is configuring the data to support the model and report higher than real R^2. For example: "Community income has been found strongly correlated with tenth-grade MCAS mathematics test scores and associated more than 80 percent of the variance in school averaged 1999 scores for a sample of Boston-area communities." (Bolon Section 3, Part B, "Conclusions.") From Table 7, below, one can see that he would have had to report "associated more than 47 percent of the variance . . . " from his original sample.

[13] When my 5-factor model is run on 2000 data, all variables have probability .006 or less. On 1998 data, the "voc" variable has a probability .064, perlimit .041, and all other variables .002 or less. Would Bolon delete "voc" from only the 1998 model for its "insignificance?" I hope not.

[14] Bolon, just before Section 2, Part C, "Observations."

[15] Section 2, part D, opening sentences.

About the Author

Stephan Michelson

Stephan Michelson is co-founder and President of Longbranch Research Associates (LRA). Formed in Washington, DC in 1979, now with offices in Maryland, North Carolina and Oregon, LRA provides statistical analysis in litigation. Dr. Michelson has been on the faculties of Reed College and the Harvard Graduate School of Education, and had minor associations with Stanford University (summer) and University of Maryland (Adjunct Professor). He has been on the staffs of The Brookings Institution, Center for Law and Education, Center for Community Economic Development and The Urban Institute (Senior Fellow).

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Spanish Language Editorial Board