REGRESSION ANALYSIS IN THE WILMINGTON CASE

JAN DE LEEUW
WITH THE ASSISTENCE OF
VIVIAN LEW, STAN BENTOW, MATT MCKEEVER

ABSTRACT. Regression analysis is defined, and its uses and abuses are discussed briefly. Our definition is somewhat broader than the usual definition, which means that it includes various forms of tabular analysis. We discuss the standard regression paradigm, which is a language to speak about effects of variables on other variables, and we argue that it has many shortcomings in large scale educational studies, especially if it is used for inference or causal attribution. Applications of formal and informal regression analyses by the expert witnesses for the defense are discussed, and some additional regression and logistic regression analyses are presented.

CONTENTS

Date. June 12, 1995.

CONTENTS

1. Introduction

Regression analysis is a very popular technique in sociology, psychology, education, economics, political science, law, and in various related disciplines. It is usually applied in a completely mechanical way, following certain rigorous recipes, which are implemented in standard computer packages such as SAS, SPSS, or BMDP. In this report I argue that regression analysis is rather *poorly understood*, because the basic principles are hidden under the purely technological aspects of the techniques. Also, the standard recipes give a false sense of security.

Moreover, regression analysis is often *misused*, in the sense that it is used to perform tasks that it is not really intended for. This does not mean that regression analysis can not or should not be applied in such situations. It is not our business to set up prohibitions. What it does mean is that researchers who use regression analysis with the standard recipe in these non-standard situations leave themselves open to serious attacks, and to a whole series of methodological objections which cannot really be answered convincingly. As long as they work in an environment in which the standard recipe is accepted without questioning, this is no problem. As soon as they move to an adversarial environment, in which the recipe itself is also questioned, this makes them vulnerable, and usually easy targets.

I discuss the basic principles and uses of regression analysis, and some of the more common abuses. The discussion is as non-technical as possible, but it also tries to avoid introducing misleading simplifications and analogies.

The report is tailored to the case of the *Coalition to Save our Children vs State Board of Education of the State of Delaware*, but of course the arguments are perfectly general. I draw on my general experience in educational and other forms of applied statistics [1].

2. Uses of Regression Analysis

Regression analysis is used to study how the distribution of an *output variable* varies in different groups of individuals. The groups of individuals are usually defined by one or more *input variables*. The output variable, of which there is only one in a particular regression

[1] Jan de Leeuw is Director, UCLA Statistics Program; Director, UCLA Statistical Consulting; Corresponding Member, Royal Netherlands Academy of Sciences; Senior Fellow, National Institute of Statistical Sciences; Editor, Journal of Educational Statistics; Former President, Psychometric Society.

analysis, is also called the *criterion* or *predictand* or *outcome* or *dependent variable* or *regressand*. The input variables are called *predictors* or *independent variables* or *regressors* or *design variables*.

In *simple regression* we have one input and one output variable, while in *multiple regression* there is also one output variable, but more than one input variable.

In the definition of regression analysis it is often emphasized that there is *only one* output variable. The important aspect is, however, that we are interested in the variation of the output as a function of the input. It is not that the output variable happens to be one-dimensional. In fact, in *multivariate regression models* we consider the variation of several outputs simultaneously, still as a function of the input.

2.1. Some Jurisprudence. The definition given above is close to the standard one. We could give hundreds of quotations to this effect, but we think a single one will suffice.

> Regression analysis, as it is presented in this article, is an important and general statistical tool. It is applicable to situations in which one observed variable has an expected value that is assumed to be a function of other variables; the function usually has a specified form with unspecified parameters. *E.J. Williams, Article on Linear Model (Regression), International Encyclopedia of Statistics.*

There is a subtle difference between this definition and the way I use the word regression. In Williams' definition the emphasis is on the *expected value* or *mean* of the output, as a function of the input. This is too narrow and too specialized for my taste. I think the term regression can be applied to the study of the *distribution* of the output variable as a function of the input variables, and thus to any *statistic* derived from that distribution (such as the mean, the variance, the histogram, the frequency count, and so on). If we emphasize the mean, we limit the term regression analysis to numerical output variables, although studies with ordinal or nominal output are also very common. Except for this relatively minor difference in emphasis, my terminology is again standard. This is illustrated by the following quotation.

> In the regression relations discussed in this article only one variable is regarded as random; the others are either fixed by the investigator (where experimental control is possible) or selected in some way from among the possible values. The relation between the expected value of the random variable (called the dependent variable, the predictand, or the regressand) and the nonrandom variables (called regression

> variables, independent variables, predictors, or regressors) is known as the regression relation. *E.J. Williams, Article on Linear Model (Regression), International Encyclopedia of Statistics.*

2.2. Describing Population Differences. Consider the following situation. We study the distribution of a variable, such as SAT score, in four school districts. This means that we can make either a table or a histogram four each of these four districts, and we can compare these four statistics.

In a slightly more complicated situation, we can introduce the year of the study as another *factor*. If there are 12 years and 4 districts, we have $12 \times 4 = 48$ groups, and consequently also 48 rows in our table, or 48 histograms. With more input variables, the size of the table, or the number of histograms, can rapidly become unmanageable. For *Coalition vs Board* UCLA Statistical Consulting has produced tables with hundreds of pages each. It is quite beyond the capabilities of human information processing to effectively deal with tables of this size.

Thus we have to apply *data reduction*. One way of doing data reduction is not looking at the whole table, or at all the histograms, but just to look at the means (observe that for binary variables, the means are proportions). This has some serious dangers. In the first place means are far from robust. In small samples they are quite unreliable, and sensitive to outlying observations. Second, means summarize only a small proportion of the actual information in the data. We throw way an enormous amount of information, and what we throw away may include all interesting effects.

The standard regression paradigm assures us that we do not really lose information by just looking at the means. It assumes that the distributions are all exactly the same, except for the means. Thus they have precisely the same shape, they are merely shifted along the axis. Even more optimistically, the standard paradigm assures us that the distributions are normal, which means that all the inetresting information is in the means and the variances (and all variances are the same). If this is actually the case, we do not throw away information at all. But in educational surveys with observational data, the standard paradigm is almost always much too optimistic.

And even if we believe the standard paradigm, we may still get into trouble. If there are a lot of predictors, we are still haunted by the *curse of dimensionality*. Ten predictors, with five values each, means about ten million populations, i.e. about ten million means. Too much

for the human mind. And not only that, to compute ten million means we need at least ten million observations. Too much for the human budget.

If there are not enough data, the model takes over. A stronger and more restrictive model require fewer data. We move to the right on the scale from empiricism (left) to rationalism (right). We assume the effects are linear and additive, i.e. the effect of a particular race and SES and achievement combination is a weighted sum of the numerical achievement score, a numerical race score (this could be just 0 for white and 1 for black), and a numerical SES score. In stead of ten million means, we only have to remember ten weights, i.e. ten regression coefficients. They contain all the information needed to compute the means.

Again, the standard paradigm may not necessarily give us good representations of the actual means, certainly not if we try to approximate ten million means with only ten parameters. But remember, we do not actually have ten million means, we only have about, say, 1,000 observations. This means we have to approximate at the most 1,000 very ill-determined means, with most means based on just a single observation. The fact we are doing a lousy job will be tend to be hidden by the paucity of our data. The model has to compensate for so many missing observations, that it becomes almost impossible to falsify. And because models which cannot be falsified are not really useful as models, this means that regression analysis in these sparse cases mainly serves as a descriptive device, a compact summary of a large number of data points, a smoothing of irregular patterns of small-sample means.

2.3. Prediction Rules. Regression analysis can also be used to construct prediction rules. The terminology *predictor* and *predictant* already suggest such a connection between regression analysis and prediction. And indeed, suppose a client walks into my office with scores of his daughter on SAT. I also know the SES and the race of the client. My job is to predict how well the daughter will do in college, for instance in terms of GPA after one or two years.

If I have a regression equation in my files with GPA as outcome and SAT, SES, gender, and race as predictors, then I can plug the daughter into the equation, and I can tell the parent what my prediction of the GPA is.

An alternative procedure, which requires a much larger file cabinet, is to select from my files all individuals with the same race, gender, income, SAT and CTBS score. Those individuals have been followed for some time, and I also know their GPA. Thus I can compute the mean

from my files, which gives me the best prediction of the daughters GPA, and I can compute the variance, which gives me a degree of confidence for my prediction.

Both procedures are based on regression, the first one assumes linearity, the second one does not. The question with these rules is not if they are "true" in some sense or another, but if they work and work good enough to keep me in the prediction business.

This then is another way to present regression results. The regression equation does not describe the actual state of the world, or the causal mechanisms underlying educational success, but it is just the best tool for prediction that the educational scientist is able to come up with (for the given price). If I tell you what a regression equation is, I am just telling you what I will do when a client walks into my office, and so on.

3. Strengths of Regression Analysis

3.1. Marginals are not Enough. If we have a number of predictors, and all these predictors are suppose to have some relationship with a criterion such as achievement, or suspension, or enrollment in special education, then some form of regression analysis is necessary to disentangle the relationships between the variables. It does not necessarily have to be linear regression, or even formal regression, but relationships between an input and the output must be studied with some sort of control for other input variables.

For example, if we are interested in the relationship between SES and Race as inputs and Suspension as output, then it can be misleading to cross SES and Suspension in a table and Race and Suspension in another table, and to present these two tables as separate findings. In order to get a complete picture of the relationships we also need the table of SES and Race, and in this particular case even the full three dimensional table is necessary. Thus we need the SES $\times$ Race table for suspended persons and the SES $\times$ Race table for non-suspended persons.

3.2. Fighting the Curse of Dimensionality. Ten predictors, with 5 possible values each, lead to 10 million profiles, i.e, combination of values. We cannot introduce parameters for each profile and estimate them, and we cannot make convincing descriptions of ten million distributions. We need data reduction.

Data reduction by only looking at selected univariate and bivariate marginals, or by cross tables of the criterion with each of the predictors,

can be misleading. It turns out that in many cases we get a considerable data reduction by looking at all bivariate tables. There are 45 of such tables in our example, and each table has 25 counts, which means we have about 1,000 numbers to look at now. In the case of approximately normally distrbiuted data we can reduce the number of parameters we have to look at at five per table (two means, two variances, and a correlation coefficient), and thus there are only about 250 numbers left.

It is clear how this type of data reduction works. We look at selected marginals, and we try to reduce further by looking at statistics which we assume show the interesting variability. The basic concept of regression analysis, which is to select subtables such that there is one outcome and a few predictirs, seems to serve us well in this respect.

4. Abuses of Regression Analysis

4.1. Inference. Often regression analysis is used to *infer* from the sample to the population. Inference is based on the notion that the data are some sort of random sample from a population, and that we use the statistics computed from the data to make probability statements about the parameters of the population. In the standard regression paradigm the errors in the regression model (the part that cannot be explained by the predictors) are assumed to be independent and identically distributed. In most case the additional assumption is made that the data are samples from normal distributions.

Several comments are in order here. In many studies, and certainly in *Coalition vs Board* the data are not a random sample, they are complete, or almost complete. We study the *population* of all students. The logic of statistical inference does not apply, or only applies in the trivial sense that all observed differences are significant.

The assumptions on which the usual significance tests are based, even if the tests would apply, are highly suspect in these educational contexts. Normality is rare, even for standardized tests in selected populations. But more importantly, the assumptions of independent and identically distributed disturbances in regression models can be justified logically only if we assume that all relevant predictors are in the system. Moreover the predictors must have been measured without error. This assumption is, in most educational surveys, impossible to defend. We often have poor indicators of ability, achievement, SES, and we have not measured a myriad of factors that could possibly also be relevant to educational achievement. Thus assuming that the

stochastic assumptions of the regression model, on which confidence intervals and significance tests are based, apply is merely self-delusion.

Again it follows that we are forced to emphasize the descriptive interpretation of regression analysis. We do not isolate basic mechnism, we just describe relationships between marginals.

4.2. Causal Analysis. Reports describing the results of regression analysis typically use causal terms. Variance is *explained*. We study in how far achievement is *due* to SES or race. This causal terminology is justified only in the context of the regression model. It must be seen as a way to talk about regression results, which is not necessarily related to the use of causal terminology in other fields, for instance in fields that use controlled experimentation.

What, indeed, do we know if the residual variance in achievement predicted from SES is 30% and predicted from race and SES is 25%. It seems that only 5% is "due to race". But this is nonsense. First, regression on race alone could easily give a residual variance of 30%. Second, SES could be just race plus measurement error, in which case predicting from SES is the same as predicting from race. The fact that our variables are *labeled* in a particular way does not exhaust their meaning.

There is a gigantic literature on causation in connection with regression and regression-related models. Some researchers feel that responsible causal attribution is possible after careful analysis of regression results, at least if there is enough external information available. The more dominant position is still, however, "*No Causation without Experimentation*", where experimentation refers to designed experiments in which there is a considerable amount of control.

Causal interpretation of the results of regression analysis of observational data is a risky business. The responsibility rests entirely on the shoulders of the researcher, because the shoulders of the statistical technique cannot carry such strong inferences.

5. Data Analysis for the Defense

We now use the general methodological discussion above to review some of the points in the expert reports written for the defense by Armor, Achilles, Walberg, Rossel, Raffel and Reschly. We shall not comment on the use and selction of data in the expert reports, tempting as it may be in some cases, but only on methodological points directly connected with regression analysis and drawing causal inferences.

5.1. Armor. The first expert report we discuss is "*Evaluation of Desegregation in New Castle County*" of David Armor.

5.1.1. *Educational Outcomes.* Armor is analyzing the observed difference in achievement scores (measured by the SAT) between black and white students. That is, in an informal way, he is analyzing the regression of achievement on race. He finds considerable differences in achievement, about 75% of a standard deviation.

> The most critical issue here is whether these achievement differences are caused by past or present discriminatory acts of New Castle County school districts, such as the former aggregated school system, or whether they are caused by other factors beyond the control of these districts.

Armor maintains that achievement differences are caused by differences in SES, and he given three arguments in this section. Before discussing them critically, we enumerate them.

(1) The black-white achievement gap nationally is between 60% and 80% of a standard deviation, according to NAEP.
(2) All four districts maintain extraordinary levels of desegregation.
(3) There are major differences between black and white students with respect to SES.

This is an informal regression analysis, but as such it does not seem to be very convincing. Obviously 75% is at the high end of typical NAEP differences. It is possible that the districts with up to 80% of a standard deviation in differences are those which discriminate *even worse* than the four New Castle districts. Moreover, the argument does not rule out the possibility that vestiges of segregation exist nationwide. The argument that all four districts have desegregated to an "extraordinary" degree seems to beg the question this trial is about. And finally, perhaps most importantly, establishing SES differences between black and white families does not in any way disprove the claim that the school system discriminates. This is because it does not rule out the possibility that there is discrimination (segregation, tracking) on the basis of SES, which of course will work mainly to the disadvantage of black students.

5.1.2. *Achievement Gap Analysis.* In this section Armor does formal regression analysis to establish his claim that race does not *influence* achievement if we control for SES. I shall try to show that the section commits some or all of the sins we have discussed above. In a later section of the report I will present a more careful version of Armor's regression analysis.

- The usual paradigm allows Armor to switch to causal terminology directly as soon as regression enters the picture. We can now suddenly "estimate the degree of the racial difference due to SES factors". As we have seen above, this is only true within the language of the regression model. The word *due* should not be given a causal interpretation.
- A variety of multiple regression models were tested. The one with the best predictions was used. It is unclear what is meant by "the best prediction". If this just size of multiple correlation, then of course the model with the largest number of predictors is the best. It would also be interesting to have a record of the regression models that were investigated but not used.
- The "gap analysis" consists in predicting achievement from SES indicators, and then finding the black-white gap in the predictions. This predicted gap is then presented as a percentage of the actual gap, and we find something in the order of 80%-90%. But what is this supposed to prove ? If achievement can be predicted very well from SES we will find a high percentage here. But we also find a high percentage if race can be predicted very well from SES.
- It would be at least as interesting to look at the various SES gaps if we first predict achievement from race. Or to look at the regression of race-corrected achievement on race-corrected SES. I have not done these analysis, but they would obviously be equally valid and equally interesting. If we remove the influence of SES, we also remove a large part of the influence of race automatically, because race is correlated with SES. This is illustrated in the figure below.
- Armor's analysis is dictated by the hypothesis that if one "corrects" achievement for SES there are no significant race differences. But the reverse hypothesis, which obviously does not correspond to his a priori's, should also be investigated. As should be the hypothesis that race has a significant effect in the joint regression of achievement on race and SES.
- Another hypothesis, equally plausible, and equally uninvestigated, is that achievement depends on SES, but different for different races. This means separate regressions for black and white students.
- Armor then goes on to include first grade achievement score as a proxy for early family effects. Even more of the black-white gap is explained. This is a clear example of the *naming fallacy.*

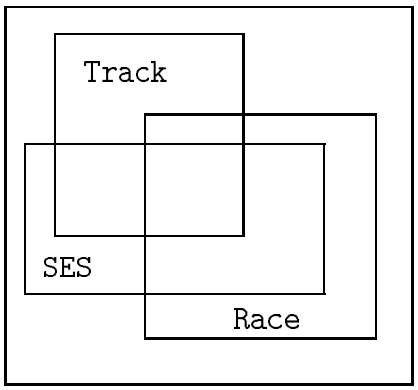

FIGURE 1. Three correlated variables

> Just because we have renamed first grade achievement, and are calling it from now on "cumulative family effects", this does not mean that it suddenly becomes conceptually identical to this. First grade achievement score differences could be due to discrimination in Kindergarten, to early and informal tracking, to differences in student abilities, or whatever. Again, using this analysis begs the discrimination question in a major way, obviously in an attempt to explain away as much of the black-white gap as possible.

It is obvious that the conclusion

> ... that the observed differences between black and white achievement derive not from school programs or policies but rather from the socioeconomic conditions found in their families and neighborhoods.

does not derive from the data analysis done by Armor, but from his personal a priori's and prejudices. What he shows is that if one controls for what he calls SES (which includes gender, by the way) then the remaining black-white gap is not large any more. Since SES is, to a large extent, "determined" by race (one should really say covarying with race), this is none too surprising.

5.1.3. *Dropout Rates.* Again, largely the same arguments apply. Controlling for SES, absenteism, and 8th grade achievement makes the effect of race on dropout non-significant. But controlling for those variables means, to a large extent, controlling for race. The fact that Colonial looks better means just that: Colonial looks better. It could be better managed, there could be less discrimination, there could even be

discrimination against whites there. This does not say anything about the other three districts.

5.2. Walberg. There is no formal regression analysis in the Walberg report, which has the title "*Academic Achievement in the Brandywine, Christina, Colonial and Red Clay School Districts*". Nevertheless there is a summary of much regression based literature. This review to me seems to be biased, and not based on scientific reasoning as I understand it.

Most of the critical factors affecting learning are beyond the control of the school. Not surprisingly, Walberg cites the first Coleman report, which has been effectively criticized in hundreds of publications. Both the data, the techniques, and the conclusions of the Coleman report have been challenged so many times, that it is already biased to mention it as a leading and authorative study. Also, obviously, the fact that school effects are not found does not mean they do not exist. Its merely means they cannot be separated by regression models from individual or family effects, which are entered at an earlier stage, and which are generally easier to operationalize and measure.

On the basis of extensive meta-analysis, Walberg has isolated nine educational productivity factor. It must be emphasized that this particular meta-analysis was criticized extensively in a recent discussion issue of *Review of Educational Research*. Walberg says that most of his nine educational productivity factors are beyond the school's control. It seems to me that this is partly a rhetorical trick, because use of the word "control" implies something absolute. Factors 4-9 can be influenced by the school, and perhaps even 2-3 are not beyond the school's reach. This is not to say that large urban public schools actually influence these factors, it merely say that they can in principle influence them (given sufficient resources).

Another ancient rhetorical trick is reflected in the title of this section, which could equally well be called "Some of the Critical Factors Affecting Learning are under Control of the Schools".

The Robinson and Branden explained variance figure of 89% is misleading, because it is based on state-level data, with corresponding inflated correlation coefficients.

The fact that "50% of a person's adult intellect is predictable by age 4 – long before school begins – and 80% by age 8" is based on a narrow and by now untenable definition of intellect. Moreover it is blatantly untrue, as volumes and volumes of more recent research into the cognitive structure of intelligence have shown. The Bloom quotation is simply an example of a discredited view of intelligence.

Moreover, even if it was true, it could equally well be used to show how stifling discriminatory tracking practices throughout the school career really are. The fact that schools do not have an influence does not mean they can not have an influence. It can mean that schools simply use mechanisms to perpetuate and strengthen the inequalities that already exist. This, of course, is not a new theory.

The last sentence of this section erodes Walberg's own thesis. Schools "cannot erase the gap", where gap of course is defined in an average sense (average over all students). Maybe schools can erase the gap for some students, maybe they can shorten the gap for all students. And this is enough reason to take school effects seriously, certainly given the fact that they tend to be easier to manipulate than basic inequalities in families or neighborhoods.

Achievement differences cannot be assumed to stem from racial dicrimination. What does this statement mean ? Not completely ? Not at all ? Not significantly ?

Disparties between black and white students exist nationally. One could argue that discrimination is nation-wide. Walberg then says, that this would imply discrimination in favor of Asian-Americans. No, this does not follow at all. If the black-white gap comes from discrimination, it does not follow that the White-Asian gap comes from discrimination too. And secondly, well, maybe Asian-Americans are discriminated positively. What is so strange about that ?

The Average Effect of Desegregation is Insignificant. Walberg's own review of the literature shows this is false. Pettigrew (predictedly) says desegregation has effect, Coleman (predictedly) says it does not. So what ? Krol reviewed 55 studies of desegregation. Only (sic!) 61% showed benificial effects. Crain and Mahard found 54% of their studies indicating a positive effect of school integration. Walberg says that is close to a coin flip – which only makes sense if he assumes that 46% of the studies actually showed a *negative* effect of integration. Presumably most of the 46% are actually null-effects.

It may be true that desegregation and integration, as operationalized, have a small effect in regression based studies of achievement. But in all cases one needs to know what the operationalizations were, what the controls were, and (unfortunately) what motivated the social or eduational scientist to write the report.

SES is a Determinant of Achievement. The causal terminology here is quite unwarranted. SES covaries with Achievement. Or, even more precisely, some variables pretending to measure socio-economic status have nonzero correlation with some variables pretending to measure

school achievement.

Factors Adversely Affecting Learning are more Prevalent in Poor Families. This is one of these findings that make social scientists look so silly sometimes. I think this fact was established quite convincingly and eloquently by Henry George, Friedrich Engels, Charles Dickens and others, ostensibly without the use of regression analysis. Again further discussion makes Walberg invalidate his own points: increasing the time children are in a supervised and stimulating environment helps. Thus schools can make a difference.

Studies of Relative Achievement Consistently Recognize the Impact of SES. It must perhaps be emphasized, as is quite common in sociology, that SES is an outdated concept, and that the idea that the social economic status of a person or a family can be measured by using a few proxies riddled with measurement error does not make much sense. It fits into the rigid, technological, static, psychometric idea that a unit (such as a student, or a mother, or a family) can be characterized by a small number of stable quantitative measures. It is true that if people are put in situations where change is impossible, then indeed change does not occur, and the world is stable and predictable. If there is dynamics and interaction, usually only available for the privileged, then change is observed. In the second section of his report Walberg uses NAEP data to study the Race Gap in achievement in the four districts, and to compare this gap with the national average. Other experts also use comparisons with the national average as an argument against discriminatory practices. On general methological grounds, this is not an appropriate argument, however. This case is about mechanisms within the four districts, not about the ultimate result of these mechanisms. Looking at marginal output-tables cannot possibly show which mechanisms are operating in the districts, and it is well known and easily illustrated that discriminatory mechanisms can lead to seemingly equitable results (and the other way around). In the Berkeley Graduate Admission case, for instance, the data showed that UCB was admitting a lower percentage of female applicants than of male applicants. This seemed to indicate discrimination. On the other hand if we looked at each major separately, it became clear that all majors were actually admitting a higher percentage of the female applicants. The key to the riddle is simple: male applicants were applying more often to the easier majors, which are admitting high percentages of applicants, and thus pooling over majors seemed to indicate discrimination in favor of males. Marginals don't tell the story about mechanisms, one needs more extensive data, and preferably longitudinal data to find out more about

what is going on. In the analysis section of this report, we make a first attempt to bring longitudinal aspects into the Race Gap problem.

5.3. Achilles. In the report "*Delaware Desegregation Case: Student Discipline Analysis*" Charles Achilles tries to minimize another black-white gap, using basically the same methodology as Armor and Walberg. Armor and Walberg maintain that the achievement gap does not exist because of discriminatinatory practices by the schools, but it is explained by the low SES of the students with poor achievement. In the same way Achilles argues that the disciplinary gap, the fact that black students are suspended much more often, should not be explained by disciplinary practices of schools but by the behavior of the students that are disciplined.

> However, when the Districts' suspension data are compared with external data sets and analyzed on a finer basis, we find that the source of difference in the suspension data is accounted for by the behavior of the students, not by the behavior of the administrators.

The basic methodological techniques by Achilles to demonstrate what he sets out to demonstrate are biased causal attributions, improper comparison, and number juggling. The basic thesis, that if one corrects for inappropriate behavior, then race no longer has an effect, is not directly demonstrated with data. A little thought also indicates that it would be extraordinarily silly to actually try to do this. If one corrects suspensions for inappropriate behavior, not much variance will be left at all, because inappropriate behavior is necessary for suspension (although certainly not sufficient).

As in the Walberg and Armor cases, the evidence is mostly circumstantial. National averages indicate high suspension indices for blacks, thus the norm for non-discrimination should be that blacks are suspended twice as often as whites, because *that* is the national average. If everybody steals, and I do not steal more than the average, then I am not a thief.

Black youths in Delaware are also arrested more, and put more in juvenile training schools, than white youths. This proves they behave worse, and are consequently equitably suspended in public schools in the districts. It may indeed be true that similar discriminatory mechanisms work in both sets of data, and one could actually use Achilles' argument in exactly the reverse direction. The arrest data prove that the administrators in state run institutions discriminate against black youths, both in the schools, in the police stations, and in the court

rooms. Of course, from a logical point of view, this argument would be as incoherent as the one Achilles makes.

Finally Achilles shows that there are other factors related to suspension indices, some of them even stronger than race. Poverty, gender, class standing are examples. We learn that "Gender transcends Race". Again, as above, this does not prove anything. These variables all covary, in some cases quite strongly. According to Achilles, gender comparisons are important, because

> .. no one attributes the higher suspension index for males to sexism on the part of female teachers.

This comparison makes no sense to me. *Coalition vs State Board* is not about racism, but about discrimination. Maybe males are discriminated against in matters of discipline. I have no idea, and neither has Achilles.

On a technical point, one does not get a good idea about the relationship between gender, race, district, class ranking, poverty, and suspension by looking at cross tables between each of the variables and suspension ratios. In the first place all the predictors will be correlated, which makes the use of separate comparisons misleading. One can argue that a strong relationship between SES and suspension weakens the case for discrimination in the school system, one can also argue that it strengthens is. It depends on the relationships between more than just two variables. Also, the use of ratios of proportions can be quite problematical, in the sense that they vary on a somewhat unfamiliar scale. Moreover, if the notion of sampling makes any sense, the standard errors of the suspension ratios will be wildly different. It makes much more sense to do a (logistic) regression of suspension on the various predictors. A number of such regressions are presented below.

5.4. Raffel. In "*Measuring the Difficulty of the Educational Task Among Blacks and Whites in the Desegregation Area*" Raffel again uses Black-White gaps (in this case ratio's). The report starts with references to the first Coleman report and to the book by Jencks, and then cites the 90% variance result of Robinson and Branden (mentioned above when discussing Walberg). We must emphasize, as we did above, that correlations on aggregated data (in this case to the state level) cannot be compared with individual-level correlations. This has certainly been known since Robinson's papers on aggregation fallacies in the 1950's.

Raffel's thesis is that educating blacks is difficult in the area served by the four districts, because many of the factor influencing educational achievement have large black-white gaps, larger than the national gaps.

He looks at all indicators separately, and does not make a single attempt to talk about their relationship, or to talk about the causal mechanisms behind those gaps. The fact that the socio-economic situation for blacks in New Castle County is worse than for whites can be interpreted as a cause of poor school performance, but it can equally well be interpreted as a consequence of poor school performance. Just indicating that these differences exist does not show much.

5.5. Reschly. Reschly has contributed "*Analysis of Minority Participation in Special Education in the Brandywine, Christina, Colonial and Red Clay School districts.*" The arguments are basically the same as the ones offered by Armor and Achilles. Indeed, there is a gap between blacks and whites in the districts, and indeed blacks are much worse off. But this is consistent with national trends, and it can be explained by patterns of poverty. If we control for poverty, then there is no meaningful contribution of race to special education enrollment any more.

In order to study the equity of the procedures followed by the districts, Reschly selected 240 students with learning disabilities at random, from each district there we 30 black students and 30 white students. These 240 case files were studied, and a large number of rather strong conclusions were drawn. I have looked at the 240 cases (actually, I could find only 230 in the file). There are 190 variables describing these 240 students, and nothing statistical can be done with a sample of this size.

If we look at the tables in the report provided by Reschly, we see that 60-70% of all minority students are enrolled in the free lunch program, and 70-80% of the free lunch students are minorities. This means, basically, that in any regression analysis poverty (as measured by free lunch participation) and race will be very highly correlated. Thus removing one of the two will remove the other almost completely, in a joint regression analysis the distribution of weights over these two variables will be very unstable, and in the relationship with other variables the two can be used almost interchangeably. It will not be possible to distinguish effects of race from those of "poverty", because the two variables are operationally equivalent (they only differ in label).

One can read Reschly's tables 6 and 7 as showing no systematic differences in enrollment with respect to race, after correction for free lunch. One can also read them as still showing an effect of race, even after the Procrustean correction for free lunch, which removes almost all race effects. It is somewhat difficult to judge the tables precisely, because no absolute numbers are given, only conditional percentages.

The appendices are supposed to give more information, but they are just reems of computer output, without table numbers and variable names.

5.6. Rossell. One of the first thing a statistician notes in Rossell's "*School and Classroom Desegregation in the New Castle County Desegregation Area*" is that the formulas for the indices used (footnotes on p. 5, 14, 15) make no sense. I have to guess what was actually meant here. I am assuming that index D on page 5 is

$$D = \frac{1}{2}\sum_{i=1}^{n} |\frac{W_i}{W} - \frac{B_i}{B}|.$$

This is just the ℓ_1 distance between the distribution of blacks over the schools and the distribution of whites over the schools. From the statistical point of view, it would make more sense to use a Hellinger or Chi-Square distance. Also, Rossell says that the index is 1.00 for perfect racial imbalance. This is somewhat imprecise. The index is 1.00 if and only if each school is either 100% Black or 100% White. Generally, I think the index may be useful for descriptive purposes, but some indication of its sensitivity and variation would be necessary to evaluate its usefulness.

The "interracial exposure index" on page 14-15 is even more mysterious. The formula for the percentage of whites in the average black child's classroom is

$$\frac{\sum_{i=1}^{n} B_i \frac{W_i}{N_i}}{\sum_{i=1}^{n} B_i},$$

where B_i and W_i are the number of blacks and whites in classroom i, $N_i = B_i + W_i$, and n is the number of classrooms. This seems to be different from the formula in the footnote on page 15, although I can't be sure because this formula is difficult to decipher.

6. Some simple regression analyses

In this section we shall present some regression analyses. The guiding principles for our selection of analyses is that they address the same questions as addressed by the defendant's experts, but hopefully in a more thorough and careful way, without relying or causal terminology and vacuous significance testing.

Another principle we have is that the analysis is done on the basis of tables, and that these tables themselves are interesting to look at. The tables are an integral part of the analysis, not something to be hidden in an appendix because it is impossible to make sense out of it anyway. We collect the tables in an accompanying document, called the "tables-document" from now on, and we collect various programs and intermediate results in yet another document, the "results-document".

The regression analyses must be seen as short-hand summaries of the tables, which are the primary source of information. Because tables are generally too large too comprehend completely, and because tables invite to "data snooping", the regression analysis is useful to summarize the most important effects in the table. To put it differently, it is an attempt to *smooth* the table, to isolate the interesting structural effects from the more fleeting accidental effects.

We have looked, in particular, at the analyses in the reports of Armor, Achilles, and Reschly. These analyses all follow the same basic pattern. They start by noticing a considerable Black-White Gap in, respectivily, achievement, suspensions, and enrollment in special education. They then set out to prove that the Race Gap becomes considerably smaller if we correct for SES, i.e. if we look at Black-White Gaps within the categories of a SES variable. We criticize their analysis by showing the following.

(1) The idea of correcting for SES, and the subsequent interpretation of the corrected gap, is fundamentally flawed. If we correct for SES, we also correct for the portion of Race that is related to SES, because of the high correlation between the two.
(2) The analysis is biased, in the sense that it attributes the maximum amount of variation to SES, and it only leaves the residual for Race. We can also start in the reverse order, correct for Race, and show that this reduces the SES Gap considerably.
(3) Even if the Race Gap is corrected for SES, it is still substantial.
(4) Additive and linear analysis of SES and Race interactions is definitely too simple.
(5) Presenting marginal tables only, which shows Race versus Suspensions and SES versus Suspensions and Gender versus suspensions

only, for instance, is highly misleading, because it never can show the mechanisms behind the relations.

(6) National comparisons with respect to output of the system are suggestive at most, because such comparisons merely show similar output, which does not imply similar relationships between the variables.

(7) Summaries of educational research on this topic are again suggestive at most. Educational research varies with times and fashions, and recent attempst at meta-analytic summaries are far from convincing. What is interesting here are the actual data for the desegregation area.

6.1. Armor, improved.

6.1.1. *Executive Summary.* We disentangle the relationship between Race and SES and achievement. Our main findings are the following.

(1) SES reduces the Race Gap by 40% for mathematics and by 50% for reading. Conversely, Race reduces the SES gap by 40% for mathematics and by 30% for reading.

(2) From the tables there is considerable evidence for interaction between Race and SES. Race has less of an effect on test scores for Poor Students, and SES has less of an effect for Blacks.

(3) The Race Gap remains substantial after correction for SES.

(4) SES has a stronger relationship with Reading than with Math.

(5) SES indicators used by Armor as so highly correlated with Race that they SES simply cannot be separated, either conceptually or statistically, from RACE.

6.1.2. *Tables.* Eight tables are presented in the tables-document. The output variables for our first set of analyses, are the raw IOWA total test scores for Mathematics (first four tables) and Reading (final four tables). All data are from 1993, and each of the four districts in the segregation area has its own table. The input variables are thus District, Race (Black, White, Other), and SES (four levels, combination of AFDC yes/no, and Free Lunch yes/no). The tables list number of elements in each of the $3 \times 4 = 12$ cells, but also the mean IOWA score in each cell, and the within-cell IOWA variances.

	AFDC Lunch	No AFDC Lunch	No AFDC No Lunch	Total
Black	305	343	426	372
White	409	457	605	590
Total	322	379	573	523

TABLE 1. Math in Brandywine, Averages

We can organize the means in each of these tables in a small 3×4 table, of Race by SES. This is done for Mathematics in Brandywine below. Actually, we only present a 2×3 table, because Other is left out from the Race values (it is a small and non-homogeneous), and AFDC - No Lunch is left out of the SES variable (it does not occur). The table shows us a Race Gap of $590 - 372 = 218$, and a SES Gap of $573 - 322 = 251$. The Race Gap is different in the different SES categories: it is 104 for low-SES, 114 for Middle-SES, and 179 for high-SES. The SES Gap is different for Races: it is 121 for Blacks and 196 for Whites.

The fact that the gaps are different for different subsets is known in statistics as *interaction*. The regression analysis we perform below assumes there is no interaction, i.e. the Race Gap is the same for all SES categories, and the SES Gap is the same for all Racde categories. This will enable us to find an average Race Gap corrected for SES, and an average SES gap corrected for Race.

6.1.3. *Analysis and Results.* The informal analysis of the tables above cannot be used to compute average effects, and is generally somewhat preliminary, because theoretically we must take the variances into account, and weight for the standard errors of the means. If a cell (i.e. a Race-SES combination) has very few observations, it should not influence the mean a great deal. But even if the cell has many observations, but a very large variance, it should not influence the mean much either. This is because if students in the cell differ widely in their test results, then the cell mean is a very poor summary of what actually goes on in that cell. Thus we should not count it too heavily.

The gap information from a weighted least squares analysis for all districts is given in Tables 2 and 3 below. Observe that the Brandywine number are a bit different from the ones we gave earlier, because of the weighting. Table 2 has the Race and SES Gaps, the Race Gap after SES correction, and the SES gap after Race correction. To study Race Gaps, before and after correction, one compares columns one and three. To study SES gaps, one compares columns two and four.

		RACE	SES	RACE NO SES	SES NO RACE
Brandywine	MATH	221	248	172	153
Red Clay	MATH	201	231	114	157
Christina	MATH	167	214	107	151
Colonial	MATH	121	148	85	105
Brandywine	READ	215	224	136	135
Red Clay	READ	194	246	92	186
Christina	READ	163	220	93	160
Colonial	READ	107	171	58	142

TABLE 2. Achievement Gaps

Table 3 has the ratio of variance due to regression to the total variance of the cell means. This is often called the variance *explained by* Race, by SES, or by Race and SES. Observe that this is not the variance accounted for in all the individual scores in a district, but the variance accounted for in the 12 means (appropriately weighted), for instance the twelve means in Table 1. Column one shows how much variance we account for if we assume that Race is the only factor, and there is no effect of SES. Column two shows the same, if we assume that SES is the only factor and there is no effect of Race. And column three, finally, show how much we account for if we allow for both SES and Race effects, but we do not allow for interaction. Thus the Race Gap is assumed to be the same for all SES levels, and the SES Gap is assumed to be the same for all Races.

		RACE	SES	RACE SES
Brandywine	MATH	.79	.70	.98
Red Clay	MATH	.71	.82	.96
Christina	MATH	.69	.75	.98
Colonial	MATH	.71	.70	.98
Brandywine	READ	.74	.77	.97
Red Clay	READ	.66	.87	.96
Christina	READ	.66	.83	.98
Colonial	READ	.50	.86	.98

TABLE 3. Achievement Predictions

Although we can see from the tables in the tables-document that there are systematic interactions between Race and SES, they are not

large in terms of explained variance. Nevertheless, it is useful to conclude that the data indicate that Race has less of an effect on test scores for Poor kids, and SES has less of an effect for Blacks.

It is impossible to get the same type of detailed information from a simple linear regression analysis, even if it comes in the form of Achievement Gap analysis. It is also easy to go too far in the direction of obtaining detailed information, by trying to control for too many variables. This produces tables which are too large, and cells counts which are too small. The conclusion from Tables 2 and 3 is that SES reduces the Race Gap by 40% for mathematics and by 50% for reading. Conversely, Race reduces the SES gap by 40% for mathematics and by 30% for reading.

6.2. Achilles, improved.

6.2.1. *Executive Summary.* Our main findings:

(1) Race is a better predictor of suspensions than SES (Lunch), and that moreover the Race Gap is more stable than SES if other variables are added.
(2) If achievement is added as a predictor of supensions, then the Race Gap goes down, but of course the causal order of achievement and suspension is far from clear.
(3) The Race Gap in suspensions is smaller at low SES levels, and the SES Gap is smaller for Blacks.

6.2.2. *Tables.* We are going to set up a regression analysis similar to the one in the previous section, but with output whether a student was suspended. Again we use 1993, and we analyze all four districts separately. The tables-document has two sets of tables.

In the first set of four tables show the relation between Race, SES, and proportion suspended (any suspension). Thus we can see that of the 301 children in Brandywine, 1993, who were black and in both AFDC and Lunch, 10% got suspended at least once. If the 56 white children in the same category, only 2% got suspended. In comparing percentage, it make more sense to look at ratio's instead of differences. Thus the Race Gap for this comparison is $\frac{.10}{.02} = 5$. For No AFCD - No Lunch it is $\frac{.10}{.04} = 2.5$. We see that for blacks there is no SES Gap, while for whites the SES Gap is 0.5, i.e. in the unexpected direction. In Red Clay the SES Gap is 1.33 for blacks, and 2.5 for whites, while the Race Gap is 1.6 for low SES and 2.5 for high SES.

Achilles also considers Gender and Achievement as predictors of suspension. Although this will tend to make the analysis far more compli-

cated, let us do the same in setting up the regression analysis. Our five predictors are race (B-W-H-O), math (Iowa, four quartiles), reading (Iowa, for quartiles), SES (free lunch), and gender. Thus there are a total of $4 \times 4 \times 4 \times 2 \times 2 = 256$ possible *profiles*, i.e. with these five variables we can describe a total number of 256 different students.

The 256 profiles are given in the next set of four tables in the tables-document. For these tables the outcome variable is the proportion of out-of-school suspensions. Again the first important point is to make the tables themselves interesting enough for detailed perusal. The tables are somewhat large, they take about four pages each, and they take some getting used to. For each of the four districts we have frequencies for the 256 profiles, plus number of suspensions for students with this profile. Thus we can look directly at various gaps. In district 31, for instance, the first four profiles are

R	M	R	L	G	S	N	p
1	1	1	0	0	3	44	0.06818
1	1	1	0	1	10	57	0.17544
1	1	1	1	0	5	142	0.03521
1	1	1	1	1	12	156	0.07692

All four profiles correspond with Blacks which are in the first quartile of both achievement tests (in Brandywine in 1993). Only the free lunch and the gender variable vary here. We see that the Gender Gap is about $\frac{.175}{.068} = 2.57$ for those not in the free lunch program, and about $\frac{.077}{.035} = 2.20$ for those in the free lunch program. The SES Gap is about $\frac{.068}{.035} = 1.94$ for girls and about $\frac{.175}{.077} = 2.27$ for boys. Although the full table is interesting to look at, it is much too rich. Too many gaps can be computed, and it is not at all clear that they will all point in the same direction.

6.2.3. *Analysis and Results.* We have to use regression to find some form of average gap. In the previous (Armor) set of analysis we looked a bit at interaction between Race and SES. Here there are too many variables to look at interaction, and we use an additive model. Also, in the previous analysis the outcome was a mean test score, while here it is a proportion. This means that we could use linear regression analysis before, but we now have to use logistic regression analysis.

The difference between linear regression analysis and logistic regression analysis is not very important for our purposes. It is mainly

technical. In linear regression analysis we predict a mean, in logistic regression analysis we predict a percentage. Because percentages are bounded between zero and one, they generally have quite different distributional properties from means. Thus slightly different (and slightly more complicated) statistical techniques are necessary to analyze proportions. The actual output of the logistic regression analysis is given in the result-document.

Two tables summarize the main results. The first table is in terms of predictions. The logistic regression predicts that certain children will be suspended, and others will not be suspended. We can compare these predictions with the actual data on suspensions. In some case the prediction and the data will be in the same direction (concorcant), in other cases they will be discordant, and in a number of cases it is not clear if they are concordant or discordant. In the table we see predictions from four models, in the four districts. Model one, coded RG, just has Race and Gender as regressors, while LG has Lunch and Gender. Model three uses Race, Lunch, and Gender, and Model four adds the two achievement variables (IOWA Math and Reading). We see that adding Lunch to the set of predictors does not do much, Race is a much better predictor than Lunch. Adding Achievement also makes a difference, although perhaps not as much as expected. The next table analyzes the gap again, in the form in which logistic regression presents it. For each of the models and each of the districts we look at the difference between the regression coefficients for Black and White to find the Race Gap, for Males and Females to find the Gender Gap, and so on. The gaps are dependent, of course, on the other variables in the model. We see the Race Gap is about twice as big as the gender gap, and also more stable, in the sense that it cannot be made to go away by adding other variables, even by adding achievement. The Gender Gap is stable as well, and about the same size as the Race Gap. As usual, we see the larger gaps in Brandywine and Red Clay, and the smaller gaps in Christina and Colonial.

District	Model	Concordant	Discordant
Brandywine	RG	63.2%	16.3%
	LG	53.1%	21.0%
	RLG	65.7%	19.1%
	RLGA	78.0%	20.0%
Red Clay	RG	63.9%	18.1%
	LG	58.6%	18.2%
	RLG	68.9%	20.2%
	RLGA	75.4%	22.5%
Christina	RG	51.8%	24.3%
	LG	46.3%	26.1%
	RLG	55.5%	27.5%
	RLGA	69.6%	28.4%
Colonial	RG	49.7%	26.7%
	LG	46.7%	27.4%
	RLG	54.7%	31.4%
	RLGA	62.7%	35.1%

TABLE 4. Suspension Predictions

District	Model	Race	Gender	SES	Math	Read
Brandywine	RG	1.52	1.03	-	-	-
	LG	-	1.01	0.76	-	-
	RLG	1.51	1.03	0.01	-	-
	RLGA	1.13	1.04	-0.14	0.35	1.38
Red Clay	RG	1.45	1.10	-	-	-
	LG	-	1.10	1.23	-	-
	RLG	1.08	1.12	0.67	-	-
	RLGA	0.90	1.12	0.53	0.19	0.98
Christina	RG	0.79	0.63	-	-	-
	LG	-	0.61	0.51	-	-
	RLG	0.71	0.63	0.16	-	-
	RLGA	0.44	0.60	-0.13	0.76	1.19
Colonial	RG	0.52	0.69	-	-	-
	LG	-	0.68	0.23	-	-
	RLG	0.50	0.69	0.05	-	-
	RLGA	0.40	0.70	-0.04	0.52	0.63

TABLE 5. Suspension Gaps

6.3. Reschly, improved.

6.3.1. *Executive Summary.* We come to the following conclusions:
(1) Enrollment in special education because of handicaps EM, LD, SE, and TM is hard to predict from Race and SES alone.
(2) The SES Gap is larger than the Race Gap for Special Ed.
(3) The effect of SES on the Race Gap is larger than that of Race on the SES Gap.
(4) In both Brandywine and Red Clay the Race Gap is still considerable, even after correcting for SES.

6.3.2. *Tables.* In the tables-document we see tables of a by now familiar form. Race and SES (Lunch with AFDC combined) are used to predict enrollment in one of the four handicap categories EM, LD, SE, and TM. The dependent variable is proportion of students in a Race-SES combination which are in special education because of one of these handicaps, either part-time or full time. Again the tables show the familiar gaps, but for special education the SES Gap corrected for Race is generally larger than the Race Gap corrected for SES. Both corrected Gaps are still considerable, however.

6.3.3. *Analysis and Results.* We complete the gap analysis again by using a logistic regression to smooth the estimates and produce best average gap estimates. We fits the models which predict on the basis of race alone, on the basis of SES alone, and on the basis of both SES and Race.

District	Model	Concordant	Discordant
Brandywine	R	41.4%	14.4%
	L	47.9%	10.8%
	RL	59.5%	17.2%
Red Clay	R	42.3%	18.9%
	L	47.1%	15.0%
	RL	58.1%	22.0%
Christina	R	34.1%	18.8%
	L	39.9%	14.9%
	RL	50.7%	22.3%
Colonial	R	32.1%	20.3%
	L	43.0%	16.8%
	RL	52.5%	24.7%

TABLE 6. Special Ed Predictions

6.4. Walberg, improved.

District	Model	Race	SES
Brandywine	R	1.11	-
	L	-	1.84
	RL	0.45	1.58
Red Clay	R	1.00	-
	L	-	1.45
	RL	0.35	1.29
Christina	R	0.64	-
	L	-	1.25
	RL	0.14	1.18
Colonial	R	0.49	-
	L	-	1.18
	RL	0.07	1.15

TABLE 7. Special Ed Gaps

6.4.1. *Executive Summary.* We have the following conclusions with respect to the development of cognitive race gaps.

(1) Black students start first grade with a relatively small disadvantage. The Gap is about 30% of a standard deviation.
(2) In Christina and Colonial, Blacks even start out with about 30% of a standard deviation advantage (a standard deviation is about 150-200 raw score points).
(3) The gap widens very quickly between 1st and 3rd grade, to about 70% of a standard deviation.
(4) It then increases through 5th and 8th grade, slowly, to about one standard deviation, and from 8th to 10th grade it does not seem to increase anymore.
(5) Gaps for reading and mathematics are of about the same size, and they grow in roughly the same way.

6.4.2. *Tables.* In the report by Walberg standardized Race Gaps are given, based on NAEP results. We would like to repeat this type of analysis on the database. It is also of interest to see what happens to the Race Gap during the school career of the student. Presumably the home background is relatively constant during that period, and changes in the gap can be plausibly argued to be due to policies and decisions of the schools. There is nothing longitudinal in the Walberg tables. Armor uses first grade scores to predict eight grade scores, but he uses first grade test scores as an indicator for SES.

Ideally, we would like to take a cohort of tenth grade students, and

follow them back in time to record all their previous test results. Eventually, we come to first grade, and we discover if the gap has widened or not. While following the cohort back in time, we would perhaps also like to keep track of background variables such as Gender and SES (AFDC/Lunch).

The analysis presented in this short note is preliminary, because we did not have the time to select a truly longitudinal cohort, with measurements at grades 1, 3, 5, 8, and 10. Instead, we perform an analysis with two time points, but in different grades. Thus the second time point is always in 1993, and the data are the IOWA Reading and Math averages of Black and White students in the four districts. The first time point is two to three years earlier, and we measure the SAT averages for the same students, and then compare the gap. We do this for four groups of students: those in 10th grade in 1993, those in 8th grade in 1993, those in 5th grade in 1993, and those in 3rd grade in 1993.

For each of the age groups, districts, and tests, we can compare the Black-White Gap at the first (usually 1991) time point and and at the second (1993) time point. The raw averages are in eight tables in the tables-document (together with corresponding frequencies and standard deviations).

6.4.3. *Analysis and Results.* We summarize the information in two tables below. The first table has Race Gaps expressed in raw score differences. This can be somewhat misleading, because we will tend to compare different tests with different ranges. Thus we also give standardized Race Gaps, which are essentially t-statistics, i.e. differences of the means divided by the corresponding pooled standard deviation.

If we look at Walberg's Table 2 we also see a general tendency for the gap to increase, with the exception of 10th grade. Because every cell of the table is based on different students, the picture is less clear than from the IOWA/SAT data we analyzed.

District	$1 \to 3$	$3 \to 5$	$5 \to 8$	$8 \to 10$
Brandywine	$46 \to 129$	$212 \to 214$	$203 \to 220$	$230 \to 213$
Red Clay	$34 \to 125$	$182 \to 196$	$180 \to 187$	$161 \to 168$
Christina	$37 \to 56$	$161 \to 157$	$147 \to 152$	$182 \to 161$
Colonial	$57 \to 3$	$93 \to 116$	$121 \to 113$	$105 \to 118$

TABLE 8. Raw Reading Race Gaps

District	$1 \to 3$	$3 \to 5$	$5 \to 8$	$8 \to 10$
Brandywine	$70 \to 150$	$218 \to 237$	$180 \to 195$	$205 \to 197$
Red Clay	$78 \to 129$	$199 \to 219$	$171 \to 166$	$181 \to 186$
Christina	$32 \to 68$	$167 \to 187$	$150 \to 159$	$149 \to 171$
Colonial	$19 \to 26$	$88 \to 106$	$94 \to 128$	$102 \to 117$

TABLE 9. Raw Math Race Gaps

District	$1 \to 3$	$3 \to 5$	$5 \to 8$	$8 \to 10$
Brandywine	$0.36 \to 0.72$	$1.13 \to 1.17$	$0.98 \to 1.04$	$1.13 \to 1.09$
Red Clay	$0.26 \to 0.79$	$0.97 \to 1.10$	$0.89 \to 0.93$	$0.80 \to 0.89$
Christina	$0.28 \to 0.38$	$0.85 \to 0.91$	$0.61 \to 0.85$	$0.89 \to 0.81$
Colonial	$0.44 \to -0.02$	$0.50 \to 0.65$	$0.67 \to 0.67$	$0.78 \to 0.58$

TABLE 10. Standardized Reading Race Gaps

District	$1 \to 3$	$3 \to 5$	$5 \to 8$	$8 \to 10$
Brandywine	$0.42 \to 0.77$	$1.10 \to 1.21$	$0.86 \to 1.03$	$1.02 \to 0.99$
Red Clay	$0.65 \to 0.79$	$1.05 \to 1.16$	$0.78 \to 0.88$	$0.90 \to 1.02$
Christina	$0.22 \to 0.45$	$0.83 \to 0.98$	$0.79 \to 0.93$	$0.80 \to 0.90$
Colonial	$0.15 \to 0.16$	$0.45 \to 0.61$	$0.59 \to 0.76$	$0.56 \to 0.60$

TABLE 11. Standardized Math Race Gaps

UCLA STATISTICS PROGRAM, 8118 MATHEMATICAL SCIENCES BUILDING, UNIVERSITY OF CALIFORNIA AT LOS ANGELES
E-mail address: deleeuw@stat.ucla.edu