MANAGEMENT ANALYSIS & PLANNING, INC.

Expert Report

Moore, et al vs. The State of Alaska

Naomi Calvo

September 9, 2005

603 Fourth Street, Davis, CA 95616 ♦ (530) 753-3130 ♦ map@edconsultants.com

Opinions of Naomi Calvo
Moore, et al. vs. The State of Alaska

Report Assessment

This review assesses two reports written by Van Mueller and co-authors Terry Schultz and MaryJo Smith, submitted as part of the *Moore v. State* case. The reports are:

- *Opportunities to Achieve in Paired Alaska School Districts: A Report to Citizens for the Educational Advancement of Alaska's Children.* (1998). Terry Schultz and Van Mueller.

- *Finance, Programs, and Outcomes for the State of Alaska's Educational System.* (2004) by Nat Cole, Van Mueller, Richard Salmon, and MaryJo Smith. Specifically, this review assesses the Executive Summary and Appendices C, D, and E which were authored by Van Mueller and MaryJo Smith.

I. Review of *Opportunities to Achieve in Paired Alaska School Districts: A Report to Citizens for the Educational Advancement of Alaska's Children.* (1998). Terry Schultz and Van Mueller.

In brief, the methods used in this report are inappropriate to address the questions posed by the researchers, the matched pairs approach is used wrongly, and the conclusions are not supported by the analysis. Each of these issues is discussed below.

<u>The methods are not appropriate for the research questions.</u>

In any research study, it is imperative that the methods match the questions. In the first two pages of their study, Schultz and Mueller say they are going to research several specific questions: what is the relationship between student outcomes and student characteristics such as race and poverty level? What is the relationship between schooling inputs (e.g. staffing, curriculum, facilities, technology, management) and student outcomes? What are the "opportunities to achieve" for Alaska students in "selected" school districts?

In order to answer these questions, Schultz and Mueller choose to do what they call a qualitative comparative case study analysis. Qualitative methods are typically useful in answering questions about processes and mechanisms, the "why" questions. For instance, it would be entirely appropriate to do a qualitative study to:

1. develop a theory of why low-income students tend to do worse in school than higher-income students;
2. explore resource use and student achievement, like why School A does much better than School B, despite similar student populations and funding levels; or
3. understand the unique challenges in rural schools.

However, if we want to know the relationship between poverty and student achievement in Alaska's public schools, or whether funding is equitably distributed, or

whether students across the state (not just in a handful of districts) are receiving an adequate education and adequate funding, we would want to do a statistical analysis, not a qualitative analysis in the manner conducted here. Few of the potential research questions discussed by Shultz and Mueller on pages 1-2 are good candidates for qualitative research.

The matched pairs approach is used wrongly.

Another issue is the "matched pairs" approach the authors use. Typically researchers employ a matched pairs methodology when they want to hold some factors constant while looking at how other factors matter. They match the pairs on all the characteristics that might be important except the one they are trying to study, and then look to see how the outcome of interest differs. If they fail to match on a key variable, then the difference in outcomes could be due to that variable rather than to their variable of interest.

To illustrate, we can use a classic example from the housing discrimination literature. Researchers wanted to know if Blacks were being discriminated against in the rental market. They matched pairs of Black and White volunteer testers on key characteristics such as age, overall appearance, and income levels, and sent them out individually to try to rent the same apartment. If the pairs were indeed matched on all other relevant characteristics except race, then if the owner turned away the Black person but agreed to rent to the White person, the researchers could infer discrimination was occurring. If the match wasn't good, then this conclusion would be invalid because the refusal could be due to differences in employment status or some other factor rather than race. The key point about matched pairs analysis is that you try to control for all the variables *except* the one you are studying.

Schultz and Mueller use matched pairs, but without following the logic that makes the method powerful. They inexplicably choose districts that *differ* on key characteristics. They start with five districts of interest, selected because they are CEAAC members. For each district they chose a matching district that is around the same size and in the same general geographic vicinity. But they do not try to match any other characteristics. Instead, they intentionally select comparison districts that differ on the percent below poverty, the percent non-white, the percent of adults with high school diplomas, the graduation rate, and test scores (p.340). They do not explain why it is important to match on size and geography or why they intentionally choose districts that differ on other key factors. This is in direct contrast to how the technique is generally used and violates the logic of the approach, rendering the comparisons meaningless.

For example, Schultz and Mueller report in their conclusions that "the percent of students graduating from high school is higher in comparison districts" and that test scores are also higher in comparison districts (p.23). This is hardly surprising, given that they intentionally selected the comparison districts to be higher achieving. What they report as a finding is an artifact of their methodology.

As noted above, it is only useful to match districts if you are trying to control for possible explanatory variables. Schultz and Mueller simply document differences between five sets of Alaska school districts. It is uncertain what is being compared and why. It is unclear what if anything we can learn from this, besides the fact that there is variation. If they simply want to get a descriptive sense of variation across schools and districts, they do not need matched pairs.

We cannot generalize from a small purposive sample to the state as a whole.

The researchers state that their study design lets them "examine districts representative of overall opportunities to learn in Alaska (p. 4). However, they chose a small purposive sample, from which it is highly problematic to generalize to the state as a whole. On page 341, they state: "The identification of purposively selected comparison pairs, and the multi-site examination of adequacy and opportunities to achieve in these districts conformed to requirements for internal and external validity and mets [sic] methodological criteria for generalizability of findings."

This is untrue. Their analysis meets neither internal nor external validity requirements. Their method violates basic social science research principles. It is almost always problematic to generalize from a small qualitative study, whether or not there are comparison pairs and multiple sites. That is why qualitative methods are more suitable to questions about mechanisms and processes and theory development.

An example from their study can illustrate why it is problematic to generalize from a small sample to the state as a whole. Schultz and Mueller conclude that "rural districts have access to fewer resources for support of direct instruction than city or borough schools" (p.10). They base this statement on their qualitative impression of five REAA districts and seven city/borough districts. However, a quick analysis of state data reveals them to be in error. On average, REAA districts spend $11,971 per pupil on instruction, while city/borough districts spend an average of $8,766 (data are from audited expenditures from the Alaska State Department of Education, fiscal year 2003). Their methodology leads them to inaccurate conclusions.

Illogical conclusions are drawn.

Schultz and Mueller find that "the percentage of operating revenues from local sources is much lower in rural districts and high in comparison districts" (p.21) and say that this is part of a 'pattern of inequity'. Leaving aside that the comparison are invalid for the moment, it is unclear why they feel this is inequitable. Rural districts probably have lower fiscal capacity than urban districts, in which case it is perfectly equitable for the state to shoulder a greater burden in financing the schools. In fact, if the state did *not* pay a greater share it would be inequitable.

Similarly, they find that "revenue from impact aid is much greater in rural than in comparison districts" and that "total revenue per pupil is greater in rural than in comparison districts" (p.21). In both cases this "disparity" actually increases vertical

equity. The rural districts, as the authors point out, have greater percentages of at-risk students, and so should get additional resources. By their own definition, this is equitable.

The qualitative analysis is substandard.

In discussing qualitative research, Shultz and Mueller explain that it is valuable because it lets us "derive fruitful explanations" and develop "new theoretical integrations" (p.330). Qualitative research lets us "create understanding about the processes by which people construct meaning and to describe those meanings" (p.331). "The data collected tend to be descriptive, consisting of people's own words and word-pictures of events and activities. The presentation of findings also employs description….There is an emphasis on process—how things happen rather than whether a particular outcome was reached. There is a concern with meaning—how the various participants in the program see and understand what happened….Accounts of qualitative reports often read more like a story than a research report. ….stories have a concrete, vivid, meaningful flavor that often proves far more convincing to a reader…than pages of numbers" (p.332-333).

This is their own description of qualitative research—but where is it in their report? Where are the stories, captured in people's own words? Where is the rich vivid description? Where are the comprehensive, in-depth portraits of the districts and their communities, of the educational opportunities being offered to students? Where are the fruitful explanations and the new theoretical integrations? Instead we get pages of tables. This isn't qualitative work, even by their own standards. In a useful qualitative study, we would be introduced to the students in these schools, to their principals and teachers and communities, and we would hear stories about them, and the stories would be presented alongside an analysis that told us something new about educational opportunities in Alaska.

In addition, the quantitative part of the analysis is uninformative because it is primarily composed of purely descriptive tables comparing pairs of districts that are more or less arbitrarily matched. These data are not of much use unless we had a logical matching system or a larger sample of districts and did statistical tests to see whether differences were meaningful or due to chance.

II. Review of *Finance, Programs, and Outcomes for the State of Alaska's Educational System.* (2004) by Nat Cole, Van Mueller, Richard Salmon, and MaryJo Smith. Executive Summary and Appendices C, D, and E.

Appendix C: Curriculum Audit

In this section of the report, Mueller and Smith survey 30 Alaska school districts and report on staffing, curricular offerings, and technology. They generalize from these 30 districts to the state as a whole, although there is reason to question the representativeness of their sample. In addition, the length of the survey instrument and

the detailed nature of the questions raises issues about the accuracy of the self report. Finally, the report is descriptive in nature and does not tie the findings to either funding levels or student outcomes. Therefore, it does not appear to have direct bearing on whether funding is adequate.

Sample Representativeness

The researchers claim that the participating districts are representative of the entire state, but they do not present the evidence necessary to support this claim.

The researchers say that "the responding districts accounted for 72% of the students in Alaska" (p.C-4), but the superintendents actually reported data for a subsample of schools that was not randomly selected. So the relevant information is the number of students in reporting schools as a percent of the total, not reporting districts. That information is not provided. In addition, the fact that 72% of students were in study districts does not by itself prove that the districts are representative. Consider this analogy: Let's say we have a jar with 100 marbles, 70 green and 30 red. If we pick out the 70 green marbles and confirm that they are all green, we would not then claim that greenness was representative of the entire jar, even though we had sampled 70% of the marbles. Likewise, it is possible that the unchosen districts differ from the ones in the study in important ways. Allowing superintendents to hand pick schools selected for the study also introduces a potential source of bias.

In order to validly assess whether a sample is representative, there are standard statistical tests to use. Generally social scientists devise a representative sample through random selection. If each district is equally likely to be chosen, and the sample is large enough, then the sample will be statistically similar to the population as a whole. Of course, through random chance it is possible that the sample might be skewed, which is why we employ statistical tests to ensure this has not happened. These two mechanisms together—the random selection of districts and the tests of statistical significance—help us be confident that a sample is generalizable. If our sample is not randomly selected, then it is particularly imperative to do the right statistical tests to make sure that the sample is truly representative.

In the Mueller and Smith report, however, neither mechanism is used. The sample was not selected randomly and the appropriate statistical tests are not presented. One way to adequately assess representativeness is to run a logit regression using study participation as the dependent variable and the demographic, structural, and attainment variables simultaneously as independent variables. Another way to examine representativeness is to do a comparison of means test on the two groups. Instead of presenting one of these established methods, the authors simply assert that most of the variables were "similar" for the districts in the study and not in the study (p.C-4). They present several data tables listing means for study districts and non-study-districts, but unless they show tests of statistical significance we cannot know whether any differences are meaningful or not.

It is noteworthy that on three key measures—the percent of limited English proficient students, the percent of students eligible for free/reduced price lunch, and average composite test scores—the authors acknowledge that the study districts are *not* representative of the state as a whole. Since test scores are one of the primary outcome measures of interest, and since the report emphasizes the particular needs of at-risk students, and since the authors themselves admit that the study districts are not representative on these measures, it is hard to argue that it is valid to generalize from the study districts to the state as a whole.

Another issue is the reliability and accuracy of the survey data. Self report surveys run the risk of being inaccurate. This is especially true when the survey instrument is long, the questions time-consuming to answer, and the incentives for accurate reporting low or non-existent. In this case the survey was exceptionally long—19 pages— and required a lot of detailed information that would have to be gathered from different sources. It would have taken many hours and much effort to fill out. Age of overhead projectors? Number of calculators? Speed of computer processors? Number of children being home-schooled? The respondents could easily be excused for making guesses or leaving sections blank. The researchers did not make any effort to check on the accuracy of the reports, and they also did not report on how much data were missing. Self reported data in this case are even more suspect since the respondents had good reason to believe that their responses would be used in conjunction with litigation in which they had an interest.

In addition to the methodological issues discussed above, the authors fail to tie their curriculum audit to either funding or performance. We cannot tell from their report whether the factors that they cite have an effect on student performance, nor can we tell whether the deficiencies are a result of insufficient funds or inefficient resource allocation.

Staffing

In their analysis of staffing, Mueller and Smith report statistically significant differences in staffing by building type, municipal type, and district size. They break district size into four categories: less than 500 students, 500-1000 students, 1001-5000 students, and greater than 5000 students. Why do they turn district size into a categorical variable, instead of leaving it continuous? And how/why did they choose those categories? No explanation is given for this. Making size a categorical variable artificially constrains the data.

More importantly: why do we care that there are statistical differences in staffing by district size and grade configuration? It seems to me that we would expect there to be differences: smaller schools should need less clerical support, elementary schools might need different services than secondary schools, etc. It seems difficult to conclude anything about adequacy from this finding. In fact, it might even be evidence of efficiency.

Courses

In the Executive Summary (p.7) the researchers talk about courses that were "deemed necessary by best practices and in our professional judgment" and include (among others) debate, journalism, marine biology, psychology, sociology, and AP psychology. They do not cite any best practices literature that says it is necessary to offer these courses to provide students with an adequate education, nor do they report what percent of high schools across the country offer such specialty courses as AP psychology or marine biology. They fail to make a convincing case that such electives are an essential part of an adequate education.

The researchers go on to say that "the types of secondary programs and support services that were offered were inconsistent across buildings and districts; not all programs, activities, or support services were offered at all schools." (p.8). They do not mention that such differences may be an important part of ensuring vertical equity. Different children in different communities may have different needs, and one of the primary tenets of local control is that districts be able to deploy resources in a way that meets the needs of their particular population. In an equitable system, we would not then expect to see identical ratios, course offerings, and activities across schools.

Technology

The authors conclude that "The knowledge level and leadership for the integration of technology into the curriculum and classroom was very limited" (p. C-43). It is unclear where this conclusion comes from, since there did not appear to be any questions about this on the survey.

Community Support

The authors do not describe how the community support indicators presented in Table 13 (p.C-44) relate to the level of resources. For instance, whether a school solicits opinions from parents, has parent/teacher groups, involves parents in program planning and assessment, or utilizes Alaska Department of Education support services, could easily be due to management, administration practices, or cultural factors rather than to funding constraints.

Appendix D: Very Small Program Sites

This section of the Mueller and Smith report lists school enrollment levels by district and describes some facts about Alaska's smallest schools. The authors assert that the State has "tended to under-provide appropriate resources for very small program sites" (Executive Summary, p. 11) but do not provide any evidence to support this claim. Their analysis makes no attempt to link funding to performance, or even assess whether small schools are under-performing. In fact, our analysis shows that there is no

relationship between school size and test scores, controlling for funding and demographic factors.

Appendix E: Costing Out and Adequate Education for Alaska's Students

This section of the report does not provide enough information on either data or methods to fully evaluate the results.

As best as I can tell from the description given, here is what the researchers did:

1. Collected FTE and salary information from 71 schools in 28 different districts;
2. Determined "adequate" staffing levels based on "several standards including best practice, current Alaska practice in high performing schools, requirements of NCLB federal legislation, and state standards" (p.E-5);
3. Determined "actual staffing costs reported for FY2002" for the 71 participating schools; and
4. Somehow combined actual staffing costs with projected adequacy FTEs to determine the "total expenditure amount per elementary building for adequacy" by ADM category (<500, 501-1000, 1,001-5,000, >5000).

This raises a number of questions:

- Defining "adequacy:"
  - Where *specifically* did the "adequate" staffing levels come from for each category (is it a state mandate, an NCLB mandate, or a "best practice" and if the latter whose best practice is it)?
  - How do the suggested FTEs compare to professional judgment models that have been done in other states?
  - The report specifies staffing levels per grade or per number of regular classroom teachers. For example, "for regular classroom teachers, a minimum of 1.0 FTE teachers per grade or 1:15 ADM students" (p.E-6). This method does not make sense for smaller schools, of which there are many in Alaska. If a K-12 school enrolls 75 pupils, one teacher per grade level would equate to 13 teachers, resulting in about six students per teacher. I have never seen a costing out study recommending staffing ratios anywhere near that low. And does a school with 50 students really need a half-time art teacher dedicated to that school? If the students spent enough time doing art to keep her busy, they would probably not be devoting sufficient time to reading, writing, and math. A quick back-of-the-envelope calculation shows this. Let's say that each child receives two hours of art instruction per week (that would be quite a lot), and that the art teacher sees pupils in groups of ten. That equals 10 hours of student contact time per week (50 pupils times 2 hours per week divided by 10 pupils per class). Even considering additional time for lesson planning, that hardly seems enough to occupy a 0.5 FTE position. The "adequate" FTEs do not seem realistic for some school sizes.

- The use of actual staffing costs from a non-random subsample of districts is problematic. Teachers are typically paid according to a teacher salary schedule which is decided at the district level. Differences in salaries by school within a district would then reflect the seniority and education level of the teachers at the school, so the aggregate salary information they collected confound teacher experience, teacher education, and geographical cost differences, for a particular subset of districts.
- It appears (though it is not clear) that the researchers constructed average salaries for each district size category, but the justification for that is not apparent, especially given that the size categories appear to be arbitrary. Why are they dividing by size category and how did they choose these particular categories? How sensitive are the results to the categories chosen? In an analysis of this type it is imperative to do a sensitivity analysis to ensure that the results are not due to the particular specification used. No sensitivity analysis is presented here.
- Also, if some of the districts are oversampled (have more schools in the sample than other districts), and if salary schedules are constructed at the district level, then these districts will be weighted more heavily in the results.
- So, it is inaccurate to estimate state-level costs from this information. A database of teacher salaries exists and is available from the State Department of Education. A more valid approach would be to use these actual figures, rather than extrapolating from a non-random subset of districts to the state level. Using the four district size categories adds a lot of unnecessary error to their estimates, especially when aggregating up to the state level. More precise methods were available, but were not used.

Conclusions: These reports do not link funding to practices, or practices to achievement. Methodologies are used inappropriately, samples are not representative of the state as a whole, and conclusions are unfounded. These reports do not meet rigorous social science standards. As such, they do not provide much useful information on the adequacy of funding in Alaska schools.