AUG 10 2012

August 6, 2012

South Dakota Department of Education
Attn: Ferne Haddock
800 Governors Drive
Pierre, SD 57501

Dear Ms. Haddock:

I would like to comment on a section of the new accountability rules, specifically 22:55:02:09, which deals with the **Calculation of college and career readiness**.

It currently states that 50% of that indicator score is generated by ACT participation, with the other 50% split evenly between achievement on the math and English sub-scores of the ACT.

Students are not currently required to take the ACT, and in fact most *register on-line, taking whatever* knowledge about those taking it away from us. We will be judged on something we have little control over. What about those students planning to attend a college that requires the SAT college entrance exam? What about those students who cannot afford to take the ACT? What about those students who immediately enter the military, or world of work, or are no IEPs and don't need to take the ACT? They may be perfectly ready to enter a career, and the school will receive no credit for them.

Thanks for the opportunity to have input on the accountability plan.

Dave Peters, Superintendent

525 EAST ILLINOIS · SPEARFISH, SD 57783-2125
Phone (605) 717-1229 · Fax (605) 717-1200

ACCREDITED BY AdvancED

AUG 2 1 2012

BRANDON VALLEY SCHOOL DISTRICT 49-2
Administration Center
300 South Splitrock Boulevard, Brandon, SD 57005-1651

Phone (605) 582-2049
Business Office (605) 582-2058
Fax (605) 582-7456

David Pappone, Superintendent
Paul J. Lundberg, Business Manager
Todd R. Williams, Operations Manger

BOARD OF EDUCATION
Jay Rasmussen, President
Renee Ullom, Vice President
Sandy Klatt, Member
Sue Hegland, Member
Gregg Ode, Member

August 21, 2012

Dear South Dakota State Board of Education:

We, David Pappone, Superintendent at Brandon Valley and Jennifer Lowery, Superintendent at Tea Area Schools, are writing to comment on the proposed rules for the Public School Accountability System, Article 24:55. Mrs. Lowery is particularly qualified to provide technical input since she holds a Masters Degree in Psychometrics from the University of Nebraska-Lincoln. We have mutual concerns that the proposed high stakes system is not sufficiently robust in statistical reliability.

One major concern comes from the proposed ranking of school buildings based on School Performance Index scores. The ranking is certainly high stakes in that the ranking will be made public and the public will make judgments about relative quality of schools based on the ranking. Those judgments will be made assuming that the SPI scores are accurate reflections of quality when, in fact, the SPI scores are not sufficiently reliable to justify these judgments. In fact, the SPI ranking and score, which indicates school quality, will have different levels of error associated with it depending on size of the group. Statistically, the error associated with the size of the gap and non-gap group is too variable to reliably measure actual school quality.

The error in the scores is best likened to sampling error in an experimental group even though the whole approach would not meet the standards necessary for a true experiment with randomly assigned treatment groups because our population (N) is too small, let alone a simple random sample (n). The results from the State of South Dakota as a whole are the general population and each school building is a sample of that population. The best samples would be random and large if one hopes to derive meaningful and reliable conclusions from the data. The proposed system uses neither truly random nor universally large samples, which is precisely what makes it difficult to verify as statistically valid or reliable.

Samples are affected in terms of randomness by geographical location, population size, particular industries that may be located in a community and a wide variety of other factors that can concentrate disproportional segments of various subgroups in some districts while not in others. The samples in South Dakota schools are clearly not truly random. The samples are also not all large. In fact, most of the samples will be relatively small due to the number of small schools in South Dakota. Without a random,

sufficiently large sample, the results should not be relied upon to make high stakes judgments.

To see the magnitude of this problem, one can calculate the standard error of a sample (i.e. school building) as being the standard deviation of the data set divided by the square root of the number of students. The standard error gives a measure of how accurate a given SPI might be. The higher the standard error, the less likely a particular SPI represents a "true" score. A low "n" has more room for statistical inaccuracy than a higher "n".

We do not know the calculated standard deviation for the data set, but the following examples assume an arbitrary standard deviation equal to one for illustration. The actual standard deviation would need to be used to calculate the real standard error associated with each sample, i.e. building. For example, a small high school that tests 10 juniors with a standard deviation of one would have a standard error of 0.32. The same would be generally true for graduation rate and ACT, the measure of College and Career Ready, or other data sets that rely on low "n" values. There were at least 16 school districts in South Dakota that had 10 or fewer juniors reported on the Fall Enrollment data found on the DOE website for last year. These same schools had a similar number of seniors, who affect the graduation rate and College and Career Ready factor. The small "n" affects each factor in the SPI similarly.

At the other end of the spectrum, a large high school that had 500 juniors (and a similar number of seniors) and the same standard deviation of one would have a standard error of 0.045. The standard error for the small "n" is over <u>seven times</u> the standard error for the large "n". The proposed accountability rules will rank order these schools as if this standard error was the same, or insignificantly different. The truth of the matter would be that the standard errors are drastically different with the SPI of a school with a small "n" being highly prone to error (volatility), while the SPI of a school with a large "n" would be statistically stable. It would be mere pretense to rank these schools based on SPI and to make high stakes decisions based on that ranking as if the difference in "n" has no impact.

One way for the State to verify the reliability of this proposed ranking system would be to back-test the SPI on previous years' data to see if schools with small "n" are stable or unstable from year to year over a period of years. We believe that some work on this may have been conducted, but we are not privy to the results or whether the results were analyzed for this issue. If our hypothesis is correct, small schools would dominate the top spots in each year, but it would not necessarily be the same schools since the standard error is just as likely to detract from a score as it is to benefit a score. We would also predict more volatility in the SPI of small school buildings compared to the volatility of the large school buildings. If this condition exists, it proves the point that these rankings are inaccurate representations of school quality.

To summarize, it is easy to see that a small school might occasionally have all, or nearly all, of its students rank high or graduate in any given year due to the "luck of the draw".

That same school might have the opposite occur based on just one or two families moving in or out of the district. The statistical probability of this happening with a large school is virtually zero. Yet, the State proposes to rank these schools without regard for the inherent error of the scores.

We would propose that the scores should include the confidence interval that accounts for this inherent error and that any schools with overlapping confidence intervals ought to be considered tied. This approach was used by the Board on Higher Education and Workforce (BHEW) from the National Academy of Science when it ranked research-doctorate programs throughout the United States (which was highly controversial among degree programs even with this statistical approach). It took the BHEW nearly ten years to develop its measurements. In a FAQ on why the BHEW used this methodology it said:

> **"Q: Why are the illustrative rankings given in ranges instead of a single number?**
>
> A: The committee felt strongly that assigning to each program a single number and ranking them accordingly would be misleading, since there are significant uncertainties and variability in any ranking process. Uncertainties arise from assumptions made in creating a ranking model based on quantitative data on program characteristics. Even with such a model, variability arises from numerous sources, including differences in the views among the faculty surveyed, fluctuations in data from year to year, and the error inherent in estimations from any statistical model. The ranges reflect some of this uncertainty and variability."
>
> *Taken from http://sites.nationalacademies.org/pga/resdoc/pga_051962.

We concur with the BHEW. We suggest that the State learn from the experience of others in making these high stakes rankings. The confidence interval would be a step in the right direction, but like the BHEW rankings, the process will still be open to criticism.

Another issue is that the key indicators for the accountability system are generally criterion referenced measures with predetermined standards while the ranking is a normative process that imposes a normal curve on the data. In effect, the system guarantees that some schools will "fail" no matter how good they are on the criterion referenced standards. The system will always have priority and focus schools. It is like the old "grading on the curve" mentality that used to be prevalent in education and has been discredited in a standards-based environment. We understand that this requirement is being imposed by the federal rule makers and is not something that can be easily corrected given that the negotiations at that level are already complete.

No Child Left Behind is traditionally based on a criterion referenced system. The classifications and ranking of schools on the SPI is normative and not meaningful in this context. We ask the Department to develop valid and reliable (meaningful) cut scores within each level of the SPI. When it comes to D-Step, our students are identified with a meaningful performance indicator, i.e. advanced, proficient, basic, below basic, based on cut scores set by the Department of Education from recommendations made by a team of psychometricians. Likewise, a school's SPI score, proficiency rating, should be identified

based on a meaningful cut score grounded in psychometric qualities. We understand the state needs to work with a certain percent of schools and therefore they will probably be rank ordered, but we ask that within that ranking they are given a meaningful SPI. The SPI must not be based on their percent rank, but rather on clear criterion grounded in psychometric qualities.

There is also major concern with the measure of college and career readiness. ACT should not be the lone measure of college and career readiness. The scores used in math and English are not designed to be indicators of both college and career readiness. ACT defines "college and career readiness" as: ".... the acquisition of the knowledge and skills a student needs to enroll and succeed in credit-bearing first-year courses at a postsecondary institution (such as a two- or four year college, trade school, or technical school) without the need for remediation."[1] This definition ignores those students who do not go on to additional schooling and assumes they are not "ready". Why were English and math the only two indicators used within the ACT? The sole use of the ACT as a predictor is a flawed system. We should be considering a wide variety of indicators: AP courses, AP Test Passing Rate, SAT, dual credit, honors programs and others such as student success in technical fields. Will there be a statewide implementation of the ACT? If there is a statewide implementation, who will pay for it? This indicator affects 20% of a district's SPI and is highly suspect as it applies to the historical ramifications of the Gap Group.

Another issue we have with the proposed rules is the provision of Incentive Points in **24:55:02:12.** We question how the Board arrived at five being the correct number of points to create an incentive. We believe that the accountability system itself creates all the incentive needed for districts to strive for improvement, if in fact districts need an incentive to perform at their best. These points would be better termed "bonus" points.

Our problem comes from the fact that we do not know the statistical significance of five bonus points and we believe it is important for the bonus points to have some statistical relationship to the standard deviation and that the magnitude of this relationship be decided on purpose with full knowledge of the statistical effect on the rankings. The question would be: What part of a standard deviation does the Board believe is appropriate for an incentive or bonus? Five points seems like an arbitrary number whose statistical significance is unknown. Those five points might represent a significant portion of a standard deviation or might not be significant at all. In the end, the awarding of these points affects the ranking of schools and adds even more confusion to the public judgment of comparative quality. We recommend not using incentive or bonus points to avoid confounding the results.

It is difficult to criticize the school climate indicator since no specific measure has been identified yet. However, we have reservations about how this will work and whether there is a sufficient research-based connection between perceptions of respondents to a survey and actual school quality. It would seem that this indicator would be highly susceptible to error based on one-time events in a school. For example: A year round approach would be supported by the research as an improvement to school quality, but it might not be popular. The school climate indicator would be impacted negatively while the actual quality of the school was improved. At this point we can only wait until the State settles on a specific measure and provide input at that point.

Are there alternatives to this flawed approach? We will leave it to the Board's statistical experts to suggest the statistical tools that would be available to help mitigate the differences in sizes of schools. We believe using confidence bands instead of a single score would be an important ingredient in arriving at the best approach. We also recommend identifying statistically significant cut scores on the School Performance Index by which schools are classified. We suggest the Board delay action on these proposed rules until it can get a thorough report on the statistical validity and reliability of this system from its own statistical experts. The report should include data that back-tests the proposed system over three or more years and should include recommendations for mitigating the hugely differing school sizes. We do not believe the present proposal meets the statistical standards necessary in the high stakes environment that surrounds accountability.

At the very least, if the Board feels compelled to approve these rules out of some practical necessity, then it should print highly visible disclaimers about the statistical accuracy, or inaccuracy, of the system. The general public needs to understand the statistical limitations of this approach in comparing and ranking schools. It may not end up mattering much since it is easy to draw conclusions from a rank-ordered list while missing the footnote.

We make these comments with good intention and do not do so with the intention of maligning the work of the very dedicated staff of the SD Department of Education. We appreciate the initiative to seek a waiver from the onerous NCLB requirements. Our hope is that our comments will lead to an even better, fair accountability system that provides the people of South Dakota a solid basis for making quality judgments about its schools.

Sincerely,

David Pappone, Superintendent
Brandon Valley School District

Jennifer Lowery, Superintendent
Tea Area School District

1. The definition of "college and career ready" comes from the document found here: http://www.act.org/research/policymakers/cccr11/pdf/ConditionofCollegeandCareerReadiness2011.pdf.

From: Stadick Smith, Mary
Sent: Wednesday, August 22, 2012 11:33 PM
To: Haddock, Ferne
Subject: Public comment

Ferne, this needs to go in your official public comment folder. We have made copies already for tomorrow's meeting.

Mary Stadick Smith
Director of Operations and Information
South Dakota Department of Education
(605) 773-7228
mary.stadicksmith@state.sd.us

From: Julie Mathiesen [mailto:jmathiesen@tie.net]
Sent: Wednesday, August 22, 2012 4:44 PM
To: Stadick Smith, Mary; Kirkegaard, Don; Duncan, Kelly; gfouberg@abe.midco.net; mhoyt@santel.net; Patricia.Simmons@southeasttech.edu; msn.com, rgowen; phelpsstacy@gmail.com; tsabers@muthelectric.com
Cc: Leidholt, Betty; Schopp, Melody (DOE); Rank, Bobbi
Subject: sharing input

Hello Fellow BOE Members,
I'm passing along an email I received from Mark Froke, Vermillion Supt. I'm sharing it with you with his permission.
See you all tomorrow, Julie

Hi Julie, I have some worries about the accountability proposal that's coming to the State Board tomorrow. The main issue I have at this time deals with points assigned to the teacher quality category based on a school indicating where teachers are generally rated on the Danielsen evaluation model as distinguished, proficient, basic, or below basic.

Vermillion School District implemented Danielsen last year and have found it to be an excellent evaluation system. As we implemented the system I've worked with the principals trying to encourage accuracy and limit rating inflation. The unintended consequence of this proposal will be the probability that schools will rate their teachers unrealistically high in order to garner all the points possible in this high stakes atmosphere. Consequently, the effectiveness of the Danielsen evaluation system under the proposed state accountability system will be at least compromised.

From a general perspective, given the magnitude of the accountability proposal, it would be prudent on behalf of the State Board to delay implementation of the proposal to allow more in-depth review of the proposals. School administrators have been focusing on the beginning of school and haven't had an opportunity to realistically respond to the proposed rules.

Best wishes and thanks for your consideration of this matter.

Mark