The BEACON

News from The Coalition for Excellence in Science and Math Education

Volume XX, No. 1   Queries? email Rebecca Reiss (next page) Copyright © May 2017

PRESIDENT'S MESSAGE, Dr. Jessica McCord

This Beacon sheds light on the New Mexico Public Education Department's (PED) use of a multiple measures model to evaluate teachers and to grade schools. We highly encourage you to read this if you are involved in education decision-making at the higher levels, if you are a teacher, if you are a parent, or if you are concerned about teacher evaluation methods. We believe this is compelling and provides sufficient evidence, both from an analytical sense and from a very personal sense, that will cause concern about the use of this evaluation model.

The first article, *A Vindication of the Criticism of New Mexico Public Education Department's Teacher Evaluation System*, is an analysis by CESE of a December 2016 peer-reviewed publication that includes a coauthor who is the primary architect of the current teacher evaluation and school grading systems in use today in New Mexico. You will find that this paper's conclusions do not support the use of the current NM teacher evaluation system. The next article, *How New Mexico's Teacher Evaluation System Translates to the Classroom*, is a personal account of one teacher's experience with the evaluation system; a journey that will leave you with a better understanding of the effect these policies have on teachers. The results are evocative. Then, we present one cartoon about NM graduation rates and the cut-off score that students need to achieve to graduate.. There is a long story behind this, but the cartoon has the answer to a question all New Mexicans interested in education should be asking about those graduation rates. Hint – look behind the curtain. Maybe we will address this in more detail in the future? We'll see.

One more note, CESE's annual meeting will be on June 24th, 1:30 pm at the UNM Anthropology department lecture hall (see the map on the last page). Our featured speaker is Dr. Frank Etscorn, the award-winning inventor of the nicotine patch. The critical discoveries that lead to the 1986 patent *Transcutaneous Application of Nicotine* were made while he was mentoring students as a Professor of Psychology at New Mexico Tech in Socorro. He will discuss a subject very near to our prime mission: Never Give Up on a Student. We think you will be entertained and enlightened by Dr. Etscorn.

EDITOR'S MESSAGE, Dr. Rebecca Reiss

I am honored to assume duties as the Beacon editor starting with the Beacon's 20th edition. Taking over this position from Kim Johnson is a bit intimidating; these are big shoes to fill! Since the traditional 20th anniversary gift is china we will be making 20th anniversary mugs available; but we need a new look! At our June 24th annual meeting, we will announce the details of a competition to design a new logo for CESE.   We will be asking those in the trenches (teachers and their students) to help us with this endeavor. Stay tuned!

The Beacon is published by the Coalition for Excellence in Science and Math Education (CESE). A 501(c)(3) nonprofit corporation, we are incorporated in the State of New Mexico. Visit our website at http://www.cese.org.

CESE annual dues are $25 for individual, $35 for family, and $10 for students. Please see last page for membership form.

E-mail Beacon submissions
to Editor, Rebecca Reiss, beetle@zianet.com

A Vindication of the Criticism of New Mexico Public Education Department's Teacher Evaluation System

Summary

We conclude that the current multiple-measures teacher evaluation model used by the New Mexico Public Education Department is neither valid nor equitably applied to New Mexico's teachers.

This conclusion is in large part based on extensive study of a technical paper[1] co-authored by Dr. Pete Goldschmidt, the primary architect of the current New Mexico teacher evaluation model, from the paper's own conclusions contained, therein, and from what we consider to be shortcomings of some of the paper's assumptions as well as key items not considered. The evaluations created using this model can, and almost certainly do, mischaracterize many teachers, some significantly, as well as resulting in an almost certain unequal application across the state.

The referenced study generally compared models using the top 10% and 25% of teachers, according to the database chosen for this study and, therefore, does not actually address the majority of teachers, which should be closer to the 50% level. Furthermore, even if this study did present error-free results (insufficient information is provided in the paper to address this aspect), these results do not support the use of the New Mexico method of evaluation or similar multiple measures evaluation models for any state. This is particularly important for any state that does not follow the "normal" student distribution for the nation, for such a culturally and demographically diverse state as New Mexico.

Synopsis of Applicable Results

The primary technical architect of the New Mexico teacher evaluation system currently being used is Dr. Pete Goldschmidt. He is the third author[2] on this technical paper in which the type of teacher evaluation system now used in New Mexico, plus other related multiple measures models using the same teacher database, are analyzed using the Bill and Melinda Gates Foundation produced database of teacher evaluation data (Measures of Effective Teaching – MET). New Mexico's system is specifically mentioned in this paper.

We are aware that Governor Martinez and Secretary of Education Skandera recently announced a change in one teacher evaluation item, reducing the weighting factor quoted in this paper from the 50% student standardized test score "growth" to 35%. However, this new weighting is also addressed in the paper's model comparisons. This is somewhat more complex than stated here, but our conclusions are unaffected.

One of the most important summary conclusions from the paper is as follows: "**We find that accuracy varies across models and cut-scores and that models with similar accuracy may yield different teacher classifications**." That is, using the same database and the same evaluation criteria, the evaluation models yield different results for teacher classifications. This conclusion in the technical paper leaves it unclear if New Mexico's system is accurate, or for that matter, if any similar systems using multiple measures, as currently used around the country, are appropriate for evaluating teacher effectiveness. The models, in fact, often provide different rankings for teachers, and there is no *a priori* way to tell if any of these are correct.

When speaking of the overall results in comparing models, the paper includes the following summary statement: "**individual teachers can still be classified inaccurately within model or inconsistently across models. For the three cut-scores used here, the models would yield inconsistent classifications for 14%, 12%, and 7% of the sample, respectively. Thus, crucially, even with a strong association in the aggregate, different combination models may yield divergent inferences at the level of individual teachers.**"

Also in the discussion of the result, the following statement is made: "**Although there is widespread consensus around the idea that robust teacher evaluation requires multiple measures, discussions are now centering specifically on what dimensions of teaching performance it is conceptually important and empirically viable to capture, and in turn how to measure and combine these dimensions in practice. Combining multiple fallible indicators does not automatically yield *better*, less fallible inferences, but it always results in *more complex* inferences.**" This means that the multiple measures method used for New Mexico does not necessarily yield reliable evaluations, but always adds more complicated inferences as far as interpreting the results—complex, especially in terms of telling any given teacher what to do better.

And, we have a qualifying statement at the end of the paper, (# 9), which states: **"Moreover, because we used a subset of complete data, our analyses may understate the extent of inconsistency in real settings where missing data will be prevalent; the results presented here may thus reflect an upper bound for reliability and accuracy in real policy applications.**" In other words, the conclusions of the paper may not even reflect how badly the evaluation models actually do perform.

Remember that all of the above statements and conclusions are from the paper coauthored by the chief architect of the New Mexico teacher evaluation system now used by the Public Education Department (PED). This is significant.

Though there is much more to this paper, these quoted statements capture the sense of the overall conclusions regarding the use of a New Mexico style, multiple measures (student performance/growth as measured by standardized tests, observation scores, and student evaluation surveys of teachers). Additionally, the authors are assuming that these three metrics are the only ones that should be included, at least as far as this paper study is concerned. In fact, there may very well be additional, even more important factors that one needs to draw valid conclusions as to teacher performance. And one or more of the evaluation elements used may be inappropriate. Furthermore, the study does not allow for inclusion of those factors, such as demographics and culture that may cause a teacher's evaluation to change significantly from place-to-place and situation-to-situation[3].

Finally, there is considerable disagreement in the education community as to whether or not standardized tests used as the metric for student performance and growth plus the use of value added modeling (VAM) schemes are even valid as far as teachers impact on student learning is concerned.[4] These are very important considerations for New Mexico's evaluation system, since it depends on methods, metrics, and metric manipulations now viewed as being of questionable validity by many education, psychometric, and statistical experts.

Conclusions and Recommendations

In agreement with, and based on the referenced paper reviewed, we conclude that teacher evaluation systems of multiple measures that include student performance results from standardized tests are inappropriate for making any high stakes decisions regarding New Mexico teachers' performances. We further conclude that (in accordance with the paper) even using the results to counsel teachers based on the current evaluation model results could be inappropriate and possibly even detrimental in many cases. (Note: it is reasonable to extrapolate these results to the current New Mexico ABCDF school grading system, too. However, there were no numerical analyses performed directed specifically at that topic. Still, the same logic almost certainly applies, since New Mexico teacher and school evaluation methods regarding the use of standardized tests for performance and student growth and VAMs are very similar.)

We, therefore, recommend in light of New Mexico's teacher evaluation model's own primary architect's conclusions in this paper, that the current model(s) should be terminated. Evaluation model(s) more appropriate should be adopted for New Mexico teachers and other education personnel and schools. (Note that school evaluation model changes may require legislative action, but teacher evaluation changes can be performed administratively.) We suggest that evaluation models used by industry, other school systems, and traditional professional areas be studied for best fit to New Mexico's education system by an independent group that includes teachers, administrators, business leaders, parents, and, as appropriate, education consultants who are experienced in this area or are credentialed individuals who have studied teacher and related evaluation models.

We strongly recommend that new evaluation models for New Mexico should be constructed by using inputs from the above group and consolidated by a non-partisan, smaller group selected by the Legislative Education Study Committee and the Public Education Department. If a constitutional amendment to reconstitute the old State Board of Education should pass, then that entity should be responsible for this task, still using the personnel cited above to provide relevant input.

Finally, we need to emphasize that clearly teachers and principals, etc., are unique in the job they do, but as professionals should still be evaluated to help them improve their performance or in some cases to advise them to seek another profession. That means that the elements of the education professions that are different from other professions must be considered. But there are still sufficient similarities with other professions such that the state need not start from scratch, but rather build on extant evaluation models, be it for teachers, principals, or even individual schools.

[1] Martinez, Jose Felipe, Schweig, Jonathon, and Goldschmidt, Pete "Approaches for Combining Multiple Measures of Teacher Performance: Reliability, Validity, and Implications for Evaluation Policy," *Educational Evaluation and Policy Analysis December 2016, Vol. 38, No. 4, pp. 738–756 DOI: 10.3102/0162373716666166 © 2016 AERA. http://eepa.aera.net*

[2] The author order in APA (American Psychology Association) style directions, of which this paper is assumed to use, generally denotes degree, but not necessary the specific type of participation. However, considering Dr. Goldschmidt's involvement with Value Added Modeling (VAM) and the topic matter, we assume that his contribution was fairly involved.

[3] "AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and educator Preparation Programs," Approved by AERA Council June, 2015: in particular Section titled **Limitations of the Use of VAM for Evaluation**, p2 ff

[4] Haertel, Edward H., "Reliability and Validity of Inferences about Teachers Based on Student Test Scores," *William H. Angoff Memorial lecture Series*, Copyright 2013 by the Educational Testing Service, Presented March 22, 2013 at the National Press Club, Washington D.C.

How New Mexico's Teacher Evaluation System Translates to the Classroom

Lisa Durkin

So New Mexico has joined almost every state in our union in creating and implementing a teacher evaluation system – yep! During the Race to the Top era of educational reform, teacher evaluation and school grading schemes were the latest-greatest accountability measure that was sure to fix all that ails our education system. Unfortunately, reform that revolved around educational accountability fell flat, mirrored by NAEP scores. Someone needs to send a memo to all the various departments of education, because the remnants of this failed reform initiative linger on today. It may be a good idea to put Race to the Top reform in the closet along with No Child Left Behind initiatives.

As with any occupation, not every teacher is competent in his or her profession. Teachers do need to be evaluated. There are several overarching questions involved with assessing a teacher's instructional quality. What do the results of this system mean, and how does it inform improved practices? Does this system use valid data to make an assessment? Are there unintended consequences?

To answer these questions let's take a look at my 2015-2016 Teacher Evaluation. I received an exemplary rating. The distribution metrics found in the total evaluation results were broken into a 50%/50% split between student performance scores and multiple measures. For science teachers, student scores come from the New Mexico Standards Based Assessment (NMSBA) or from End of Course (EoC) exams for a given subject if it is available. PARCC exams are used for English language arts and math teachers. Teachers of other subjects may be judged only on EoC results or other approved standardized tests with a small fraction being exempted from test result evaluations altogether.

I mostly teach a freshman Integrated Science course and one section of Astronomy. Neither Astronomy nor Integrated Science has an EoC. Only juniors take the NMSBA. For 2015-2016 there were only nine 11th grade Astronomy students who took the NMSBA, so 50% of my evaluation was based on nine student scores.

For each teacher, student scores are averaged over a three-year period, as a student growth measurement. This is to insure that each cohort's mix of students doesn't skew the data. Only 35 students over a three-year period provided growth data for my evaluation.

There were 86 biology students from my 2014-2015 evaluation who disappeared from the 2015-2016 evaluation. That is probably because they were not my students – I don't teach biology. Yet, their scores were 50% of that year's evaluation.

There are several other issues with how NMSBA scores are used to determine a teacher's evaluation results. Eleventh grade scores actually measure the contributions of three teachers, because the science NMSBA assesses material that students learn over a three-year period, not just the subject taught in the 11th grade. Furthermore, I teach Astronomy. Astronomy comprises only 20% of the material tested by the NMSBA.

Recently the PED decided to reduce the percentage that student scores weigh in calculating a teacher's evaluation. Student performance on standardized tests will account for only 35% of teacher evaluations beginning this year. That means that the other 65% will be accounted for by using multiple measures like classroom observations, attendance and student surveys. Traditionally, classroom observations were the means by which teaching quality was measured. It was a crude system with only five sections that were scored with pass or fail. The current system is far more extensive in how it assesses a teacher's competence.

In today's evaluation system, classroom observations are broken into four domains:

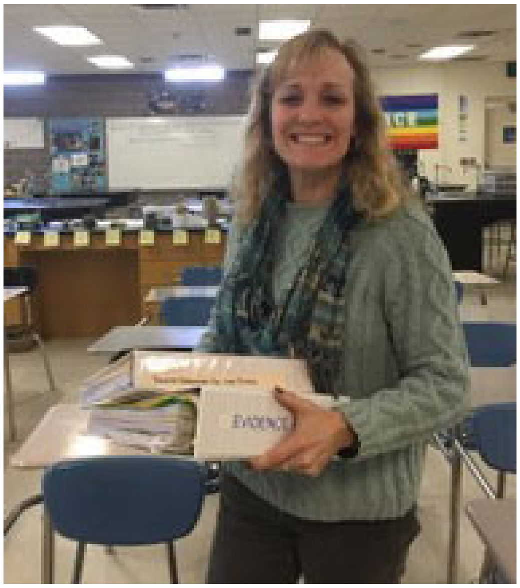
#1 Planning and Preparation

#2 Creating an Environment for Learning

#3 Teaching and Learning

#4 Professionalism

Each domain has at least five sections (A-F) broken into five levels (from ineffective to exemplary) based on criteria points for each level. Administrators begin their evaluation from level three. If all of the criteria are met, they move to level four. If a criteria point at any level is not met, the designation drops down a level. So, if for domain 1 section B, a teacher does not document all of the criteria points for level three (effective) their designation drops to level two, minimally effective. I counted the criteria points for levels three through five, for each section, for all four domains. There are 272 criteria points that a teacher must document to reach the exemplary level. For my exemplary designation I documented all 272 criteria points. **That required 710 pages of evidence.**



Exemplary teacher Lisa Durkin with binders of evidence demonstrating her competence. Was it worth her time?

My evaluation rating changed over a three-year period from Effective to Exemplary. Over that time period I used the same teaching methodology in the same subject, teaching students with the same demographics, etc. The only difference was that the third year I collected 710 pages of documentation. What does the multiple measures portion of the evaluation mean if the same teacher using the same methodology improved their rating solely because of documentation? If a teacher does poorly on a given section, is it because they are deficient or because they didn't provide sufficient evidence of competence?

It took a great deal of time generating and documenting 710 pages of evidence for my evaluation, and my administrator had to spend hours poring over my documentation to make his final determination about my classroom observation rating. It would have been far more beneficial if we could have spent that time on our students and their needs.

Many resources at school sites must be reallocated to accommodate the necessary software, data generation, computer resources, computer technician expertise, etc. for the teacher evaluation as well as the tests necessary to generate data for teacher evaluations. Counselors spend a great deal of time on test coordination when they could be working with kids.

The window for EoC testing is layered on top of the same time frame for Advanced Placement exams and directly follows PARCC and NMSBA all of which must be proctored by a teacher. Continual testing, up to three months for Juniors, consumes instructional time, which would be better spent on teaching the curriculum being tested.

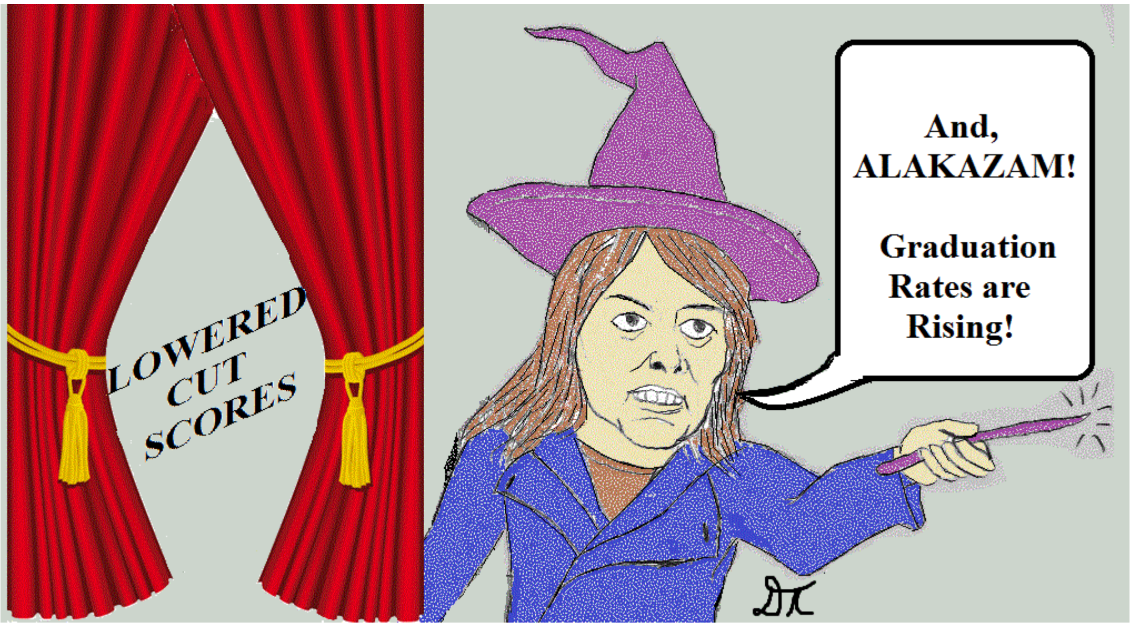
EoC tests are reconstructed every year and have contained test item mistakes. The testing platform provided by PED changes almost every year and this year's version is so rife with errors that test results could easily be considered invalid, because students could submit their tests unwittingly thinking they had answered all the questions when they hadn't. EoCs are of questionable importance except that they are used for teacher evaluations.

We all want exemplary teachers in every classroom. Teachers need to be evaluated and held accountable. An evaluation system needs to provide a clear avenue for teachers to improve their practices without undue burden on school resources that could otherwise be used for kids. Do the **results** of the evaluation system meet the **goals** of the evaluation system? Are all the unintended consequences worth it? Are test results a valid means to measure teacher accountability?

Accountability measures implemented in the name of educational reform haven't improved education. Instead, teachers are leaving. For the 2016-2017 school year, 36% of the teachers at my school site moved on. This educator exodus is not solely due to the morale-crushing evaluation system. There are certainly other stressors in the profession. Sure every job is hard, and teachers sound whiney when they complain, but if being an educator under current conditions were within the normal range of work related strife, there wouldn't be such high turnover.

This evaluation system isn't effective at making me a better teacher. It doesn't give me appropriate and useful feedback. The evaluation process is overly cumbersome, time-consuming and tedious. It robs kids of my time and attention. Data from student performance uses too small a sample size, and encumbers the time and efforts of too many teachers, to provide a valid assessment of a teacher's contribution to student achievement. There is no discernable value added to the instructional environment because of this evaluation system.

We need to step back and analyze what does work with this system and make improvements. Any evaluation system must not drown teachers in extra paperwork. Scapegoating teachers for low student performance on tests is like blaming a dog for having fleas.



NM PED: Pay No Attention to What's Going on Behind the Curtain !

Membership Dues/Donation Form
Coalition for Excellence in Science and Math Education (CESE)
501(c)(3) non-profit, tax deductible
Dues and Donations cheerfully accepted year round
(Expiration date is found on address label)
Member $25.
Family $35.
Student $10.
Lifetime: $500 Individual, $750 Family.
You may contribute through United Way, PayPal, or USPS.
Mail checks to CESE, 803 Maverick Trail SE, Albuquerque NM 87123.

New Membership [ ] Renewal [ ] (Please indicate any changes for renewing members) Donation [ ]
Name ______________________________ Date________
Profession and/or affiliation(s)______________________________
(e.g. Science teacher, member of APSD)
Mailing Address ______________________________
____
Phone ____________ Cell ____________ Fax ____________
E-mail______________________ (Most of our communication is by E-mail).

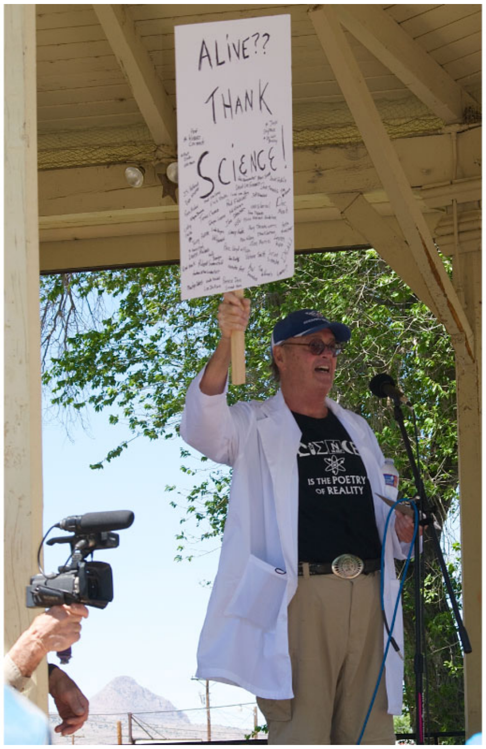
Please send change of address to Dave Thomas <nmsrdave@swcp.com>

http://www.cese.org

Coalition for Excellence in Science and Math Education
803 Maverick Trail SE
Albuquerque, NM 87123-4308

Return Service Requested

Science & Math
CESE
Education



Dr. Frank Etsorn at the 2017 Socorro March for Science

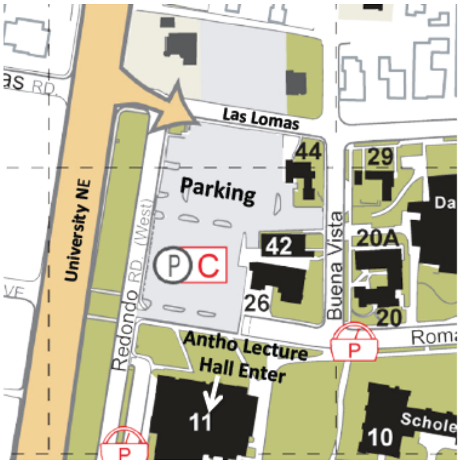
We are pleased to announce that **Dr. Frank Etscorn**, inventor of the transdermal drug delivery system known as The Patch, will be the keynote speaker for our 2017 CESE membership meeting! Dr. Etscorn's talk will be

"Never Give Up on a Student."

Saturday, June 24, 2017
1:30 PM
The UNM Anthropology
Lecture Hall

FREE and open to the public

Directions: From Central and University, go north on University until you get to Las Lomas. Turn right, then right into the parking lot. The lecture will take place in the Anthropology buliding lecture hall, immediately south of the parking lot. Parking is free on Saturdays. We look forward to seeing you there.


as RD
Las Lomas
44
29
Parking
Da
University NE
Redondo RD (West)
P C
42
20A
26
Buena Vista
20
VE
Antho Lecture
Hall Enter
Roma
P
11
Schole
10
P