Options for Incorporating Student Academic Growth as One Measure of the Effectiveness of Teachers in Tested Grades and Subjects

A Report to the North Carolina Department of Public Instruction

from

February 1, 2012

Introduction

The materials in this packet summarize findings from an investigation conducted by WestEd researchers at the request of the North Carolina Department of Public Instruction (DPI). In November 2010, we began working with the DPI to help it identify a model of student academic growth that could be incorporated into the state's existing Educator Evaluation System (NCEES). Specifically, our role was to review information about different growth models and recommend one model for adoption as the sixth standard in the evaluation of the effectiveness of North Carolina teachers in tested grades (currently 3–8 and high school) and subjects (currently English language arts and mathematics).

To date, our study has included the following steps:

(1) ***Conduct Literature Review.*** We conducted a review of research literature that targeted theoretically and empirically based support for various statistical models of student growth, including those classified as value-added models (VAMs). We sought to identify the relative strengths and limitations of the models used most widely for the purpose of estimating teacher effectiveness.

(2) ***Identify Evaluation Criteria.*** We developed a list of criteria to be considered when evaluating a growth model's appropriateness for the purpose of measuring teacher effectiveness. Criteria that were identified include the following: (a) technical adequacy (validity, reliability, and fairness) of the model for the intended purpose; (b) face validity with teachers and other stakeholders; (c) theory- or research-based support; (d) ease of use statewide for incorporating a measure of student growth in the NCEES; (e) resource requirements; and (f) policy implications.

(3) ***Review Expert Panel's Technical Report.*** The DPI commissioned a panel of experts (representatives from institutions of higher education in North Carolina with advanced training in statistics and educational measurement) to conduct analyses to help the DPI judge the technical adequacy of various models for estimating a teacher's effect. Specifically, the panel sought to (a) identify a set of statistical models with the potential to yield trustworthy information about student academic growth; (b) conduct a series of analyses, using both actual and simulated data, that allowed for comparison of the relative strengths and limitations of each model from a technical perspective; (c) draw conclusions from its findings about model appropriateness and present those conclusions to the DPI; and (d) share its findings with WestEd technical advisors to allow for professional dialogue about methods, findings, and conclusions. Based on the criteria it used, the panel identified three models as having the level of technical quality needed for the purpose of estimating teacher effect: a three-level hierarchical linear model (HLM III); a univariate Education Value-Added Assessment system (EVAAS) model developed by William Sanders (EVAAS-URM); and a student fixed effects model (SFE). We reviewed the panel's draft report for key

evidence related to evaluation criterion (a)—technical adequacy (validity, reliability, and fairness)—of the model for the intended purpose.

(4) ***Respond to Technical Report.*** Following our review of the technical report, we responded to the panel with questions, to ensure full understanding of its methodology and to explore the practical importance of model differences. Specifically, we requested more information about assumptions that must be met for each model (e.g., vertical scale; how missing data are addressed), rationale for the criteria used during decision-making, and results from grade- and content area–specific analyses. We also asked for more information about the panel's rationale for eliminating from consideration a multivariate EVAAS (EVAAS-MRM).[1]

Synthesis of Findings and Presentation of Recommendations

We have developed summary tables that synthesize all of the information collected in the steps previously described and submitted them to DPI. We also have developed a set of recommendations for consideration by the DPI. These recommendations are as follows:

- We agree with the expert panel's decision that only VAMs should be considered by the DPI for incorporation into the NCEES. VAMs use longitudinal data to estimate a student's academic achievement level at some point in the future, given his or her past performance, and assuming that that student has an average schooling experience. In a VAM, if a student outperforms his or her predicted score, that student's teacher is credited with having an instructional effect beyond what was expected (i.e., the teacher added value). Importantly, these models seek to enable a fair measure of the influence of teachers on the rate of student progress. They are supported by a strong and steadily growing body of research and are widely used in districts and states across the nation.
- The expert panel reviewed eight VAMs and concluded that three stood out as having strengths from a technical perspective, particularly in relation to predictive validity and reliability (measurement precision and rating stability). These three demonstrated the lowest levels of false identification of ineffectiveness, a critical consideration. Although each statistical model demonstrated specific limitations, each also offers unique strengths. We accept the panel's findings and thus focused our validation work on these three models.[2]
- From our perspective, among the three models identified by the expert panel, the EVAAS-URM emerged as the top contender after consideration of all evaluation criteria, including technical adequacy, research-based support, and cost. The panel found it to be a consistent high-performer statistically, its use does not require a proprietary license, and

[1] As of the time this report was submitted to the DPI, this information had not yet been provided.

[2] Few empirical studies have explored the finely grained differences among the various statistical models that are classified as VAMs. This is a highly specialized subset of growth models, and, while each model has unique strengths and limitations, all of these models are similar in important ways. The considerable work of the expert panel in designing and implementing strategies for tackling this topic was appreciated, as its recommendations provided a strong foundation for WestEd's subsequent work.

model assumptions are transparent and unlikely to be violated in the North Carolina context. If no other information about the EVAAS-MRM (see sixth bullet, below) is forthcoming, the EVAAS-URM would be a defensible first-choice model.

- The HLM III is the only model in the final pool of three that allows for estimation of the effect of teachers in grade 4. If the DPI believes it is critical to select a model that includes these teachers, the HLM III may be the model of choice. Our reservation about this decision is that we believe this model is more appropriately used for measuring school effect.
- The expert panel found that, from a technical perspective, the SFE performed as well as the two other models recommended by the panel for estimating a teacher's effect on student growth. If the DPI determines that its priorities are statistical parsimony and low cost (i.e., economy), the SFE should be considered. Our caution about its use is that the model does not provide some of the valued features (e.g., flexibility) associated with the EVAAS models.
- While engaged in the model validation process, we concluded that the DPI may want to also consider a fourth model that was not recommended by the expert panel due to concerns about its feasibility for statewide implementation: the multivariate EVAAS (EVAAS-MRM). Nonetheless, we encourage the panel to reconsider this model and determine if it is worthy of further analyses. If the panel agrees to such work and the EVAAS-MRM performs better than the EVAAS-URM from a technical perspective, this model becomes very attractive. Unquestionably, it is a complex model and is more resource-intensive than the EVAAS-URM; it also will cost more to develop and implement, requires a high level of expertise during specification, and can be challenging to interpret and use. However, this model offers benefits in terms of flexibility (capacity to adapt over time to changing needs), and it is also the only model that can account for the effect of multiple teachers on a student's annual growth. If the DPI can afford the cost associated with this model (proprietary license required) *and* if the expert panel's secondary analyses support its technical quality, the EVAAS-MRM may be a top contender.
- Regardless of the model selected, we endorse the DPI's expressed intent to use the value-added estimate as only one component of the NCEES for teachers in tested grades and subjects. As currently planned, teachers' value-added estimates will be combined with their performance on five other standards during decision-making about their levels of effectiveness.
- The DPI should plan for short- and long-term research agendas that support ongoing monitoring of model use, evaluation of emerging consequences (intended and unintended) and impact over time, and cost-benefit analyses. It is advisable that, as part of its long-term research agenda, the DPI continue to monitor findings from emerging research, such as the work being conducted in the Bill and Melinda Gates Foundation-funded Measures of Effective Teaching (MET) project.