Arts Education Policy Review
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vaep20

The Inchworm and the Nightingale: On the (Mis)use of Data in Music Teacher Evaluation

Mitchell Robinson[a]
[a] Michigan State University, East Lansing, Michigan, USA
Published online: 12 Dec 2014.

To cite this article: Mitchell Robinson (2015) The Inchworm and the Nightingale: On the (Mis)use of Data in Music Teacher Evaluation, Arts Education Policy Review, 116:1, 9-21, DOI: 10.1080/10632913.2014.944966

To link to this article: http://dx.doi.org/10.1080/10632913.2014.944966

ARTS EDUCATION POLICY REVIEW, 116: 9–21, 2015

ISSN: 1063-2913 print / 1940-4395 online
DOI: 10.1080/10632913.2014.944966

Routledge
Taylor & Francis Group

ARTICLES

The Inchworm and the Nightingale: On the (Mis)use of Data in Music Teacher Evaluation

Mitchell Robinson
Michigan State University, East Lansing, Michigan, USA

The music education profession is faced with two serious problems regarding the (mis)use of data in music teacher evaluation. The first has to do with the quality and kinds of data that music teachers have been forced to use; the second is concerned with how these data are being used in the music teacher evaluation process. The evidence I will use to support this argument will come from two sources. First, I will present a short review of policy briefs targeting the use of data in teacher evaluation in general education. Then, to provide a music-specific context, I will turn to the body of scholarship in music education that has focused on these issues. I will conclude the article by providing suggestions for how we can both use better data in music teacher evaluation and use these data in better ways to inform music teacher evaluation. I will also offer some general recommendations for consideration by music educators and policymakers interested in improving the process of music teacher evaluation.

Keywords: data, music teacher evaluation, value-added measures

The following story (Tunks 1987, 53) was a favorite of the noted music education scholar Robert Sidnell, who often used it to highlight the differences between measurement and evaluation for his graduate students:

> Peering through a tangle of reeds, a hungry robin saw a cute little green inchworm, and prepared to devour him. The inchworm, thinking quickly, pleaded for his life by offering a bargain: "Don't eat me! I'm useful. I measure things. If you spare me I will measure your gorgeous feathers." And with that, the robin snatched up the inchworm, and deposited him on her back, where the inchworm carefully measured off and reported back the results. The robin was so pleased with his new friend's resourcefulness that she carried him back to the aviary where other birds needed to be measured.
>
> One by one, the enterprising inchworm measured the parade of birds that were marched before him. He slithered down the toucan's beak; he climbed across the cardinal's wings; he belayed up the heron's leg, and rappelled down the pheasant's tail. In short order, he became the aviary's measurement expert, and provided the birds with elaborate measurements of things they didn't even know needed measuring: wings, tails, feathers, nests, eggs, branches. .. even the tiniest of baby birds.
>
> The inchworm quickly grew accustomed to a certain level of respect among the birds, and in exchange for his services he lived a very comfortable life in the aviary. No longer worried about being gobbled up for a snack, he busied himself by finding ever more obscure items to measure, and presented the results in ever more arcane ways to his appreciative audience.
>
> Then, from a distant land where inchworms were rare, a beautiful nightingale appeared above the aviary. As the songbird floated in the breeze, she observed the inchworm going about his daily business—measuring beaks, wings and nests. Growing weary of the bustle below, and feeling a bit peckish, the nightingale landed gently and approached the little worm. Smiling, the nightingale trilled a challenge to the inchworm: "Measure my song."
>
> It had been so long since the inchworm had been asked to measure something other than length, or width, or breadth, that he was flummoxed by the request. "But how can I do

Address correspondence to Mitchell Robinson, College of Music, Michigan State University, 208 Music Practice Building, East Lansing, MI 48824, USA. E-mail: mrob@msu.edu

> that? I measure things, not songs," said the inchworm. Not to be dissuaded, the nightingale replied, "Measure me, or you will be my breakfast."

In the traditional version of the fable, the inchworm is cast as the clever hero, using his skill in measuring things to stave off certain doom from his hungry captor, while the nightingale is seen as the vainglorious fool who presents our hero with an unsolvable challenge. Given recent changes in the educational landscape brought about by pressures from the corporate education reform movement, a more modern retelling of the tale might switch these roles, with the inchworm playing the role of the educational technocrat who trusts in the infallibility of data to provide solutions to the problems in public schools and the nightingale playing that of the plucky music teacher, wondering how her discipline's intricacies and subtle nuances can be measured by the inches, agates, and picas of the ruler.

At this point in the tale, the sly inchworm distracts the nightingale and quietly crawls away, having for the time being safely fooled his feathered friends with the wizardry of his measuring prowess. In the schools, the situation is not quite so clear, and the story is not yet over.

PURPOSE AND ARGUMENT

In this article, I will argue that music educators are faced with two serious problems regarding the (mis)use of data in music teacher evaluation. The first has to do with the quality and kinds of data that music teachers have been forced to use; the second is concerned with how these data are being used in the music teacher evaluation process. The evidence I will use to support this argument will come from two sources. First, I will present a short review of policy briefs targeting the use of data in teacher evaluation in general education. Then, to provide a music-specific context, I will turn to the smaller body of scholarship in music education that has focused on this issue. I will conclude the article by providing suggestions for how we can both use better data in music teacher evaluation and use these data in better ways to inform music teacher evaluation. I will also offer some general recommendations for consideration by music educators and policymakers interested in improving the process of music teacher evaluation.

EVALUATION IN (MUSIC) EDUCATION: IMPROVEMENT AND ACCOUNTABILITY

Before turning to our examination of the uses of data, it may be worthwhile to briefly discuss the role and purpose of evaluation in education. The two primary purposes of educational evaluation are to *improve* instruction and to *increase the accountability* of educational programs. While these twin goals of improvement and accountability are sometimes in conflict with one another, our current policy dialogue surrounding evaluation in public education seems to be focused on only one of these aims—accountability.

The result of uncoupling accountability from the improvement of instruction in the use of data has been that for perhaps the first time in our nation's educational history, we now see policies that are *punitive* rather than *educative* in nature. According to Hargreaves and Braun (2013):

> The last issue concerns the magnitude or severity of the consequences. A principal tenet among U.S. policy makers today is that for an educational accountability system to have the desired impact, it must result in significant consequences. This belief is at odds with much of the research in education and in other sectors [Hout and Elliott 2011; Pink, 2009; Rothstein, Jacobsen, and Wilder 2008; Springer 2009], which shows that large, extrinsic rewards can dampen intrinsic motivation and that tryouts of such reward systems yield minimal to no improvement. The belief in the necessity of significant consequences is also out of step with the accountability practices of high performing countries such as Canada, Finland and Singapore that do not attach external rewards or punitive consequences to the extremes of performance on achievement tests.

Without the leavening influence of improvement, this narrow focus on data-driven accountability has resulted in policies that emphasize punishment. For example:

- In 2010, every teacher at Central Falls High School in Rhode Island was fired because of low student test scores (Khadaroo 2010).
- In February 2014, students at a school in Tennessee were given after-school detention and extra homework assignments after scoring poorly on practice tests for a state math exam (*Tennessee Parents* 2014).
- In December 2011, New York City school officials announced that twelve schools had been placed on a school closing list, primarily because of low scores on state tests (Phillips 2011).

In each of these cases, and in the majority of similar situations nationwide, opponents of these sorts of punitive policies point to poverty as the root cause of the difficulties faced by students, teachers, and schools—and yet these policies do little to address those issues of poverty that clearly impact educational attainment (Otterman 2011).

There is sometimes a misconception that teachers of music and art (i.e., "nontested subjects") fear evaluation and accountability, and that this is part of the reason that these subjects are not currently tested. In my experience, nothing could be further from the truth. Music teachers have embraced public performance and scrutiny of their work since music first entered the public school curriculum in 1838 (Mark 2008). Concerts, recitals, and "informances"

are textbook examples of transparency and accountability in ways that standardized tests can never be. These events welcome parents and other community stakeholders into the schools to witness firsthand the transformative power of music to influence student growth in positive ways.

And yet, too many school leaders fail to understand the authenticity of these forms of assessment to capture the true essence of student learning (Robinson 1995). Far from being "afraid" of accountability, music educators, I would suggest, should be recognized as the true pioneers of accountability by virtue of the public nature of their enterprise as performers and teachers.

THE PROBLEM(S) WITH DATA

What *is* true is that music teachers are highly mistrustful of the kinds of data that are being used to determine their effectiveness as educators; specifically, they are mistrustful of the (mis)use of standardized tests and assessments of student achievement in subjects other than music—typically, math and reading in grades 3–8—in their own evaluations. U.S. Secretary of Education Arne Duncan (2009) has echoed music teachers' apprehensiveness about the misuse of these forms of data, saying, "Test scores alone should never drive evaluation, compensation or tenure decisions. That would never make sense. But to remove student achievement entirely from evaluation is illogical and indefensible."

Perhaps the most pernicious trend that drives current education reform initiatives is the singular and sole reliance on data as evidence of student learning and teacher effectiveness (Berliner and Biddle 1995). While few educators would deny that the judicious use of data from student assessments can be useful in improving teaching practices, it is the apparent privileging of data as the only form of evidence used in making school policy decisions that many teachers find troubling.

Providing further context for Secretary Duncan's comments cited previously, the National Association for Music Education recently developed a position statement on music teacher evaluation, offering the following guidelines and suggestions:

> 1. Measures of student achievement used in teacher evaluation:
>
> a. Must be based on student achievement that is directly attributable to the individual teacher, *in the subject area taught by that teacher*. Student achievement measures must be used with care, ensuring that they accurately reflect a given teacher's contributions. . . .
>
> 2. Successful Music Teacher Evaluation: . . .
>
> b. Must include *measures of music student achievement* along with the above indicators, as only one element of a teacher's evaluation. *For evaluation of music teachers, measurements of student achievement should include evaluation in the three general areas of creating, performing, and responding*. The relative weighting of measures in these three areas should be carefully designed to be commensurate with the nature of the class taught and the express educational goals for that class. . . .
>
> d. *Must avoid using school-wide measures other than those directly associated with music achievement.* If the use of school-wide measures of attendance, dropout and graduation rates, and/or work habits is mandated, they should account for a minimal part of the music teacher's evaluation. (National Association for Music Education [NAfME] 2011, emphasis added)

The clear message here is that music educators agree with Secretary Duncan that student achievement is a critical factor in determining teacher effectiveness—but that this relationship is predicated on the use of data that are directly associated with music learning and are not drawn from schoolwide tests in other subjects.

The music education profession also agrees with Secretary Duncan that removing student achievement from the equation makes little sense. Student learning in music is the goal of all comprehensive music instruction and remains at the core of what we do as music educators. However, it is useful to remember that while an excellent teacher is the most powerful school factor in determining student growth and learning, nonschool factors account for over 60 percent of the variance in student achievement, with as little as 7.4 percent of this variation being attributable to teacher quality (Hanushek, Kain, and Rivkin 1998, 21).

The educational community has been aware of this phenomenon for decades. In 1966, the Coleman Report concluded that "student background and out-of-school factors are significantly more important" (Amrein-Beardsley 2014, 85) than in-school resources in influencing educational outcomes, and organized these factors into the following categories:

- Student risk factors, such as emotional and/or learning difficulties, English-language proficiency, and racial/ethnic minority background
- Student motivation and desire to do well on tests
- Parents' attitudes toward education
- Domestic stability and support, including access to books, technology, and other resources
- Access to adequate health care and proper nutrition
- Access and exposure to arts, culture, and travel, which are correlated with families' socioeconomic backgrounds and poverty levels

More recently, the American Education Research Association and the National Academy of Education issued a policy brief identifying some of the nonschool factors that influence students' achievement as measured by standardized test scores. These factors include:

- Home and community supports or challenges
- Individual student needs and abilities, health, and attendance
- Peer culture and achievement
- Prior teachers and schooling, as well as other current teachers
- Differential summer learning loss, which especially affects low-income children
- The specific tests used, which emphasize some kinds of learning and not others, and which rarely measure achievement that is well above or below grade level (2011, 1)

So while music educators understand that the judicious use of data can contribute to effective teacher evaluation, there appear to be significant problems with the kinds of data being used in many places for these purposes. Using faulty data is one problem; compounding that error by inappropriately applying systemwide measurements to individual teachers is quite another.

VALUE-ADDED MEASURES

One popular use of data in determining teacher effectiveness is the approach known as *value-added measures*, or VAMs. VAMs have been used in school districts in many states to provide statistical evidence for the impact of individual teachers on student learning. However, this is not the purpose for which these tools were originally intended or designed.

Simply put, a VAM is a statistical model that attempts to show the differences between schools in terms of their effectiveness in promoting student learning. The tool works like this: a model is built that predicts what a student's score will be on a particular test based on that student's prior test scores and other "relevant characteristics" of both the student and the school (Hargreaves and Braun 2013, 18–19). Each student's actual score on the test is compared against the predicted score, with the difference between the scores being that student's contribution to the school's value-added estimate. The average of all students' differences is the school's VAM score. A school with many students who outscore their predicted results receives a positive value-added estimate, while schools with students who do not score as well as predicted receive a negative value-added score.

While this may seem like a logical use of data to inform educational decision-making, there are serious problems with VAMs when they are used to evaluate individual teachers. For instance, because schools, classrooms, and families are not constructed or selected randomly, the predictive "power" of the statistical model that undergirds the VAM cannot "accurately disentangle all the confounding factors in order to isolate the relative effectiveness of different schools" (Hargreaves and Braun 2013, 19). The misapplication of schoolwide data to high-stakes evaluations of individual teachers is an error that creates significant stress and trauma for many schools and teachers on a yearly basis.

Another problem arises from the volatility and variability in schools' VAM estimates from year to year. Likely caused by small sample (i.e., class) sizes and inaccurate test results, many schools that receive "above average" ratings one year will be placed on "warning lists" the following year, with the opposite scenario being just as common. As Hargreaves and Braun (2013, 19) note, "These seemingly inexplicable but highly consequential fluctuations are not only demoralizing to many school staff, but also damage the credibility of the accountability system as a whole."

Sass (2008) found that among a group of teachers from five urban school districts across the country who scored in the bottom quintile one year, less than a third had similar ratings the following year, while nearly half received the highest rating the next year. The same phenomenon was found for the "most effective" teachers, with only a small percentage receiving the highest rating the following year, and the rest moving into other rating brackets.

A third problem with using VAM results as a component of teacher evaluations is the confusion between correlational and causal forms of data. According to a recent statement issued by the American Statistical Association (ASA):

> VAMs typically measure correlation, not causation: Effects—positive or negative—attributed to a teacher may actually be caused by other factors that are not captured in the model. . . . VAMs should be viewed within the context of quality improvement, which distinguishes aspects of quality that can be attributed to the system from those that can be attributed to individual teachers, teacher preparation programs, or schools. Most VAM studies find that teachers account for about 1% to 14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions. Ranking teachers by their VAM scores can have unintended consequences that reduce quality." (2014, 2)

The ASA statement goes on to suggest:

> The quality of education is not one event but a system of many interacting components. The impact of high-stakes uses of VAMs on the education system depends not only on the statistical properties of the VAM results but on their deployment in the system, especially with regard to how various types of evidence contribute to an overall evaluation and to consequences for teachers. (6)

Recently, Polikoff and Porter (2014) studied the relationships between value-added measures of teacher performance and instructional quality by analyzing data from over 300 math and English teachers in six school districts from New York City; Dallas; Denver; Charlotte–Mecklenburg; Memphis; and Hillsborough County, Florida. The authors concluded: "Overall, the results are disappointing. Based on our obtained sample, we would conclude that there are very weak associations of content alignment with student achievement gains and no associations with the composite measure of effective teaching" (15–16).

In another recent policy brief examining the impact of reform initiatives on the perceived professionalization or deprofessionalization of teaching, Milner suggests that VAM strategies are also exerting substantial pressures on teachers and administrators to ensure positive test results, which has led to instances of widespread cheating in cities such as Atlanta and Phoenix (2013, 10). Milner points to research that has found serious statistical issues plague the reliability and validity of VAM data for purposes of high-stakes teacher evaluation, including small effect sizes and large margins of error with test scores (Mathis 2012), "significant methodological errors" in many value-added models (Amrein-Beardsley 2008; Briggs and Domingue 2011), the presence of a "ceiling effect" that does not adequately differentiate teaching quality (Ballou 2009; Koedel and Betts 2010), and "attribution error" that results in the misapplication of VAM results and "raises fundamental ethical questions about the use of value-added methods for high-stakes decision making" (Kennedy 2010).

Braun (2005) of the Educational Testing Service concluded the following from his review of research on the use of VAMs: "VAM results should not serve as the sole or principal basis for making consequential decisions about teachers. There are many pitfalls to making causal attributions of teacher effectiveness on the basis of the kinds of data available from typical school districts. We still lack sufficient understanding of how seriously the different technical problems threaten the validity of such interpretations" (17).

The preponderance of research on the use of VAMs in teacher evaluation suggests that while this sort of data may be useful at the school district level in predicting some differences among schools, it is inappropriate and invalid to use these data to determine effectiveness ratings of individual teachers because of significant technical and measurement problems. These problems are only magnified when VAM strategies are used in the evaluation of music teachers.

MUSIC TEACHER EVALUATION

Over the past few years, music education policy groups have started to focus their attention on issues surrounding music teacher evaluation. Delegates from the National Association for Music Education have met with congressional leaders and Department of Education representatives (NAfME 2012b), "held several national symposia on music assessment and teacher evaluation, released a list of recommendations for music teacher evaluation (NAfME 2012a), drafted a position statement (NAfME 2011), and released teacher evaluation workbooks" (Shaw 2014). Additional state-level and national policy groups have also been created to address these issues, including the Partnership for Music Education Policy Development in Michigan and the Music Education Policy Roundtable, an advocacy alliance sponsored by NAfME.

While the dialogue among leaders in state and national professional organizations has begun to center on issues of music teacher evaluation, the scholarly literature on the topic remains sparse. In a forthcoming publication, Aguilar and Richerme (2014) surveyed *Music Educators Journal, Art Education Journal,* the *Journal of Dance Education*, and other professional publications for papers targeting the impact of the federal Race to the Top reform initiative on music teacher evaluation. The results were scarce, perhaps indicating a disconnect between what is happening in the field and the policy discussions at the state and national levels (24).

In a recent study of music teachers in Michigan, Shaw (2014) found that nearly 95 percent of music teachers surveyed disagreed with the statement, "Student test scores in math and reading should play a part in music teacher evaluations." Shaw also found strong support among the music teachers surveyed for some form of professional "portfolio evaluation" of teachers' skills (78% agreed or strongly agreed); the belief that student musical skills can be measured accurately by performance using a teacher-designed test, rating scale, or rubric (88% agreed or strongly agreed); and the belief that student skill growth should be measured by individual musical performance (71% agreed or strongly agreed). Perhaps the most provocative findings of Shaw's study had to do with music teachers' beliefs about the role of large ensemble festival or contest ratings in the teacher evaluation process.

Festival Ratings

One way in which music educators have attempted to apply data-driven strategies to music settings is through the use of large ensemble festival ratings. Perhaps because of the pervasiveness of the ensemble competition model in school music, many music educators have considered the use of the scores generated at these festivals as data relevant to determinations of their effectiveness as teachers.

Shaw (2014) asked participants in his study a number of questions about the use of student growth measures to assess student gains in achievement over time. While the respondents' support for demonstrating growth among

their students was generally quite strong, teachers clearly prioritized data on skill growth over data on knowledge growth:

> Most teachers agreed that skill growth could be measured through a student's individual musical performance and strongly supported the use of a "teacher-designed test, rating scale, or rubric" over accompaniment software. Teachers mostly disagreed with the measurement of individual growth through ensemble performance scores and were most unified in their disagreement with using student test scores in math and reading as part of music teacher evaluations. (13–14)

Hash (2013) examined several issues related to the use of large ensemble contest ratings in teacher evaluation, focusing primarily on reliability and validity concerns. The research on ensemble contest ratings reliability suggests that multiple nonmusical factors may influence contest scores (e.g., performance order, race of performers and conductors, ensemble names or labels, judges' training and experience, familiarity with the repertoire and medium under evaluation, length of the contest day, difficulty of repertoire, ensemble size, conductor expressivity, participation of exceptional learners, types of adjudication forms used), and that the inter-rater reliability of final ratings is generally acceptable but can vary widely from one adjudication panel to the next (Hash 2012). In sum, the findings on the reliability of contest ratings are inconclusive and suggest that a host of extramusical factors play a role in determining festival scores.

In terms of validity, Hash (2013) suggests that ensemble contest ratings are not designed to measure music learning in a broad sense, but they do collect information on an "ensemble's performance of a few selections and possibly sight-reading at one point in time" (165). Although NAfME encourages music teachers to cover a wide range of learning goals identified by the National Standards in Music, including composing, improvising, arranging, and understanding music in relation to other disciplines and cultures, ensemble festivals are designed to assess student learning much more narrowly (i.e., performing, singing, and perhaps sight-reading). Hash cautions that festival scores tend to be disproportionately skewed toward the top end of the rating scales, with over 90 percent of ensembles receiving grades of "I" and "II" out of the five grade ranges offered at most festivals.

Hash also points out that "large-group festivals. .. do not measure individual skills or musicianship, and only to some extent do they assess the improvement of an ensemble from one academic year to the next as a result of differences in personnel, repertoire, and adjudicators" (2013, 165). These limitations would seem to make ensemble ratings somewhat less valid for measuring changes in individual student learning over time, the primary goal of most student growth strategies. Hash concludes:

> Contest ratings provide valid assessments of student achievement only in relation to group performance on a narrow range of tasks and a limited number of national standards. Furthermore, ratings might not adequately indicate the true quality of an ensemble's performance or measure growth from one year to the next. Nonetheless, festivals do evaluate achievement on a large and important aspect of the ensemble curriculum and may provide meaningful data in an assessment system that includes multiple and varied measures. (166)

With so many music teachers attending these contests, it is understandable that many teachers would be interested in using large ensemble festival scores as one component of their effectiveness ratings. But the unintended consequences of incorporating these data into high-stakes evaluation ratings must be carefully considered. In Michigan, the state unit of the Society for Music Teacher Education has created a policy statement on music teacher evaluation that suggests that music teachers and their administrators consider the following points:

> 1. All organizations that sponsor rated festivals should establish and periodically calculate statistical reliability (consistency) for ratings generated at these events, and provide data indicating the average rating and frequency counts for each final rating (I–V) issued within a particular classification, and for all participants combined. These data will serve as norms used to compare individual results with those of similar groups. This effort may require the assistance of college faculty or others knowledgeable in statistics and education research.
>
> 2. Festival ratings are valid to the extent that they measure an ensemble's performance of two or three selections, and sight-reading ability, at one point in time. They furthermore only provide assessment for one of the five Michigan Music Standards and related benchmarks.... A complete assessment of student growth requires multiple and varied measures of musicianship and musical understanding.
>
> 3. Teachers never should be required to attend a particular festival or use the results of these events as value-added data in their annual evaluation. Music educators who choose to use this data as part of their evaluation should do so voluntarily and as one of multiple measures of student growth.
>
> 4. Teachers, administrators, and other stakeholders in music education should be aware of the numerous factors that can influence performance adjudication. According to the extant research, these might include (a) conductor and performer appearance, (b) performance order, (c) repertoire selection, (d) adjudicator experience and background, (e)

> adjudicator knowledge of special circumstances, (f) the evaluation form, and (g) adjudication procedures. . . . All of these nonmusical factors may contribute to measurement error and reduce the statistical reliability (consistency) of the final scores. Furthermore, ratings do not account for circumstances related to instruction such as rehearsal scheduling, financial support, staffing, or other factors that might influence instruction, student achievement, and the quality of the final performance, most of which are out of the teacher's control. (Michigan Society for Music Teacher Education 2011, 4–5)

The research on the use of large ensemble contest ratings appears to suggest that while there are serious issues with the reliability and validity of these ratings—mostly due to the influences of nonmusical factors such as performance order, problems with inter-rater reliability, and the narrowness of what these contests measure with respect to the totality of the music teacher's instructional responsibilities—large ensemble festivals can still provide valuable feedback to students and teachers about their group performance. These contest ratings, however, should not be considered as valid data in the measurement of individual student learning, nor should they be used as more than a single component of a more robust, multiple-measures approach to determinations of music teacher effectiveness. In this way, ensemble contest ratings can be seen as an important tool for the improvement of instruction, but as less useful as an accountability measure.

Better Data

If we are interested in helping music teachers advocate for better evaluation procedures, then it stands to reason we should be concerned with the kinds and quality of data being used in these systems. From the previous discussion, it seems clear that music teachers are not in favor of being evaluated using assessment data collected from tests in other disciplines, and that there are serious problems associated with using the data derived from VAMs in the evaluation of individual teachers.

Nonetheless, "as of September 2013, 35 states and the District of Columbia Public Schools now require that student achievement is a significant or the most significant factor in teacher evaluations" (National Council on Teacher Quality 2013, i). The weight given to data derived from what the National Council on Teacher Quality terms "objective measures of student achievement" varies from state to state, from a high of 50 percent in Michigan to as low as 25 percent in Maine. According to Siebert (2013), "acceptable" forms of evidence for student achievement tend to fall into the following four categories:

- Standardized test scores
- State exams
- Third-party assessments
- District- or locally developed assessments

For music teachers, the final two categories appear to hold the most promise for generating useful and appropriate forms of data regarding student learning that might be applied to teacher effectiveness ratings. However, many music teachers do not have the time or the measurement expertise to develop their own assessments. Fortunately, there are groups at both the national and state levels that are engaged in initiatives designed to provide music teachers high-quality assessment tools for use in collecting appropriate and authentic evidence of student learning.

NAfME recently published two workbooks intended to "offer teachers, peer evaluators, and administrators clear structure, guidelines, and examples for professional evaluations of music teachers" (NAfME 2013). One of the workbooks is aimed at ensemble teachers while the other is targeted at general music teachers. Each workbook includes "an instruction manual, an eight-page form that principals and teachers can use for evaluation, and a series of worksheets offering rubrics in areas ranging from general knowledge to classroom management to professional development" (NAfME 2013). These resources were designed by content-area specialists in ensemble and general music teaching, and contain sample activities, lesson plans, and assessment tools that may be adapted by teachers for use in their own settings.

In Michigan, the Michigan Arts Education Instruction and Assessment (MAEIA) project, a joint initiative of the Michigan Assessment Consortium, Data Recognition Corporation, and the Michigan Department of Education, was charged with defining what high-quality arts programs look like and with developing rigorous, standards-based assessments in dance, music, theater, and visual arts. The group's work resulted in the following four tools, which are available free of charge to all arts educators:

- "The Michigan Blueprint of a Quality Arts Education Program": A goal-setting document for arts education program and school improvement purposes
- "Michigan Arts Education Blueprint Research and Recommendations": This companion document to the Blueprint provides users with supporting documentation for each criterion and indicator in each arts discipline.
- "Michigan Arts Education Program Review Tool": A self-study tool that districts and schools can use to analyze and reflect on the status of their own arts education program
- "Michigan Arts Education Assessment Specifications and Prototype Assessments": A set of recommendations for appropriate assessments in the arts based on

the Michigan Merit Curriculum and aligned to national standards (Michigan Arts Education Instruction and Assessment [MAEIA] 2013)

These tools were designed by teams of public school arts educators and collegiate arts teacher educators, with project management guidance provided by consultants from the Michigan Assessment Consortium and the Michigan Department of Education. School districts in the state have been invited to compare the data from the MAEIA Program Review Tool with the characteristics found in the accompanying Blueprint, identifying strengths and challenges present in their specific contexts. These tools were designed with the following purposes in mind:

- Support district policy as well as develop district and building practices that ensure adequate time, staff, and resources for high quality arts programming for all students.
- Support implementation of sequential arts instruction, for all students, delivered by certified arts educators.
- Support the use of assessment practices and measures, yield accurate information and ensure are communicated effectively.
- Support the sustained, discipline-based, job-embedded professional learning for staff delivering arts education.
- Inform program planning, review, and improvement. (MAEIA 2013, 8)

The MAEIA Assessment Specifications and Prototype Assessments provide an example of third-party assessment tools that teachers can use to generate data on student learning in music. These assessments have been developed in four basic forms: performance tasks, performance events, constructed-response items, and selected-response items. Performance tasks are typically multipart projects that require students to research a topic, write a paper, compose or arrange a song, create a presentation and/or a performance, and reflect on what they learned during the process of the task. Performance events may be thought of as "on demand" activities that expect students to provide a response to a prompt with little or no advance preparation or research. Constructed-response items are open-ended and require students to create their own response to a given prompt. Selected-response items are multiple-choice, true-false, or similar types of questions that present the student with a variety of responses and ask him or her to select the most appropriate choice.

These different forms of assessment provide models for music teachers to adapt and use in their classrooms and ensemble rehearsals to give their students rich learning experiences. The data gathered from these assessment activities can be used to demonstrate authentic growth in student learning over time, without resorting to activities that do not promote higher-order or critical-thinking skill development such as those involving vocabulary lists and flash cards with music terms.

Better Models

The use of better data is a good step toward improving the quality of music teacher evaluation practices, but without thoughtfully designed and implemented evaluation systems, even the best data will not lead to appropriate and authentic assessment practices. The education profession has struggled with teacher evaluation for much of its history, and music teachers in particular have long felt that the systems in place are inadequate in terms of providing meaningful feedback on their practice. Improving the system, however, is a deceptively complicated task.

Making matters worse, many of the recent "improvements" in teacher evaluation have been imported directly from the business world. While this may seem to be a logical step on its face, it assumes that schools and classrooms can be run and managed like businesses—an assumption that is not only naïve, but potentially damaging to students and teachers.

The predominant employee evaluation tool in the business world for years has gone by the name of "stack ranking," and it should sound familiar to teachers and principals everywhere. The system works by dividing employees into arbitrarily predetermined "effectiveness categories," based on ratings by managers. At Microsoft, for instance, employees were rated "on a score of one to five, with one being the best. Managers were then given a curve to base their rankings on, and forced to give a certain percentage of employees a poor 'five' label—even if the managers did not consider the employee to be unsatisfactory at their jobs" (Halleck 2013). This ranking and sorting procedure was used at the software company to determine bonus payments as well as employee terminations, with "a second consecutive low ranking [meaning] that an employee would be terminated" (Halleck 2013)—the ultimate in high-stakes assessment.

Also known in some places by the term "rank and yank," stack ranking has been used historically as a means to trim a company's workforce and is typically only in place for short periods of time. Detractors of the approach from Microsoft say that it contributed to a culture of competition among employees and was the "most destructive process" in place at the software giant (Warren 2013).

Even as similar systems are currently in place at most public schools in the United States, Microsoft, Expedia, and Adobe Systems are among a growing number of companies that have now abandoned the practice of stack ranking (Brustein 2013), with Microsoft's leaders citing concerns that the system led to "a 'lost decade' and a

cannibalistic culture" (Warren 2013). As Microsoft's head of human resources explained when it made the decision to discard stack ranking, the company would now emphasize teamwork and capacity building among employees instead of pitting workers against one another and undergoing endless cycles of performance reviews; there would be "no more curves . . . [and] no more ratings" (Warren 2013). Expedia's leadership characterized the move away from stack rankings as "an effort to rehumanize the relationship between employees and their bosses" (Brustein 2013).

The main problem with using stack-ranking systems in teacher evaluation is that the approach is predicated on a series of faulty assumptions. Proponents of this approach believe that:

- teachers are the most important factors influencing student achievement,
- effective teaching can be measured by student test scores and is devoid of context,
- large numbers of American teachers are unqualified, lazy, or simply ineffective, and
- if we can remove these individuals from the workforce, student test scores will improve. (Amrein-Beardsley 2014, 84–88)

The argument is seductively convincing. There is, however, little to no research-based evidence to support any of these assumptions. As explained previously, there is no research to support the contention that teachers are the most important influence on student achievement—outside factors account for 80 to 90 percent of the variance in student learning, while in-school factors, such as teachers, account for only 10 to 20 percent. While we devote great amounts of time and resources to haphazard attempts to improve the teaching force by increasing measurement efforts, we are ignoring the devastating influences of poverty on students' lives. Refocusing our philanthropic and policy efforts from building better teacher evaluation systems to addressing issues of child poverty would be a welcome policy change, and one that is long overdue.

Similarly, much of the rhetoric surrounding the so-called "teacher effect" not only ignores the role of poverty in student learning, but actively dismisses this concern. In an editorial for the *Washington Post* in 2011, Secretary of Education Arne Duncan claimed that "school districts and their local partners in inner cities and rural communities are overcoming poverty and family breakdown to create high-performing schools, including charters and traditional public schools. They are taking bold steps to turn around low-performing schools by investing in teachers, rebuilding school staff, lengthening the school day and changing curricula" (Duncan 2011).

Teach for America founder Wendy Kopp puts this idea even more succinctly: "Education is the tool to get kids out of poverty" (Bragdon 2013). Kopp's rhetorical distancing of poverty's impact on student learning positions her as what Thomas (2012) refers to as a "no excuses reformer," along with other leaders in the corporate reform movement such as Bill Gates, Michelle Rhee, and Arne Duncan and charter school chains such as the KIPP: Knowledge Is Power Program. Thomas (2012) explains:

> "No Excuses" Reformers insist that the source of success and failure lies in each child and each teacher, requiring only the adequate level of effort to rise out of the circumstances not of her/his making. As well, "No Excuses" Reformers remain committed to addressing poverty solely or primarily through education, viewed as an opportunity offered each child and within which . . . effort will result in success.

> "Social Context" Reformers have concluded that the source of success and failure lies primarily in the social and political forces that govern our lives. By acknowledging social privilege and inequity, Social Context Reformers are calling for education reform within a larger plan to reform social inequity—such as access to health care, food security, higher employment along with better wages and job security.

Again, the research here is abundantly clear. When results are controlled for the influences of poverty, nearly every international test of student learning shows that American students score at the top of the rankings. For example, when test scores for U..S students on the 2009 Program for International Assessment (PISA) exams were disaggregated by poverty levels, American children from middle– and upper–socioeconomic status families performed as well or better than students from the top three nations in the rankings: Canada, Finland, and South Korea (Walker 2013a).

The third assumption, that a large percentage of the teaching force is ineffective, ignores the fact that the evaluative measures used to inform stack-ranking approaches are not precise enough to accurately identify the bottom 5 or 10 percent of "bad" teachers, and that using normed or "curved" ratings only ensures that 50 percent of teachers are arbitrarily and automatically assigned the label of "ineffective" based on student test scores. In other words, "by statistical design, there will always be some teachers who will appear relatively less effective simply because they fall on the wrong side of the bell curve" (Amrein-Beardsley 2014, 87).

Finally, while few teachers would argue against removing ineffective educators from the classroom, there are real concerns about the ability of the evaluation systems in place at the present time to accurately identify poor teaching based solely on the data currently being used. There is also no evidence that ineffective teachers are replaced

automatically by more effective teachers. Furthermore, while some vacant positions might be filled by those who enter the profession via alternative routes (e.g., Teach for America), "research evidence continues to suggest that these teachers remain in teaching only for the short term (e.g., three years on average) and their effects compared to other new and career teachers are no different from average" (Amrein-Beardsley 2014, 88). The preponderance of evidence suggests that even if we could accurately identify and remove ineffective teachers from our schools, and there is little reason to believe that this is possible given current systems and assumptions, there is no guarantee that doing so would have the intended results.

Our first step, then, should be to develop a system of teacher evaluation that is based on better assumptions. I would suggest the following:

- While teachers are a critical in-school factor influencing student achievement, there are a host of nonschool factors that also play important roles in learning.
- Effective teaching requires administrative and community support, adequate resources, and an institutional commitment to continued professional growth.
- The vast majority of teachers are dedicated, hard-working, and collaborative professionals who are committed to improving their practice.
- The goal of any teacher evaluation system should be the improvement of instruction and not the provision of data for high-stakes accountability programs.

Another foundational assumption upon which music teacher evaluation should be based is the importance of making the evaluation process discipline specific. That is, music teachers should be evaluated based on their abilities as music teachers and their students' growth as musicians over time, and by those with content-area expertise. One example of music teacher evaluation that may prove of interest in the design of authentic models is Connecticut's Beginning Educator Support and Training (BEST) Program.

The BEST Program was an initiative of the Connecticut State Department of Education that was developed to help improve the quality of the teaching force in the state. The novice teachers who were participants in this program were required to demonstrate mastery of essential teaching competencies related to content knowledge, planning, instruction, and assessment. These competencies were assessed through a discipline-specific teaching portfolio submitted during the second year of teaching in which beginning teachers documented a unit of instruction built around important concepts or goals in a series of lessons, assessed student learning, and reflected on their students' learning and the quality of their own teaching. The portfolio included lesson plans, videotapes of teaching, examples of student work and student assessments, and teacher assessment and reflective commentaries. These portfolios were then assessed by groups of experienced and rigorously trained music educators and music teacher educators who were selected using recommendations from administrators and colleagues.

Although the BEST Program was designed as an evaluation system specifically for beginning teachers, the program had a number of design elements that are worth considering for use with all music teachers. First, the data used in the BEST portfolios were generated by the teacher's own students in the course of normal classroom activities; there were no standardized test results or data from assessments of subjects outside the teacher's certification area. At the core of the BEST portfolio were the videotapes of classroom teaching and assessment activities, which provided rich, authentic data that represented the candidate's skills and knowledge as a music teacher. The data in the portfolio were also of the teacher's choosing and thus reflected the goals and objectives established by each teacher for her or his classroom setting.

Second, the BEST portfolios were evaluated by the novice teachers' peers: experienced, practicing music teachers and music teacher educators who were familiar with the teaching settings and contexts in the state. This kind of content-area match is critical for ensuring that music educators are evaluated accurately and appropriately.

Third, the BEST portfolio encouraged music teachers to plan for, teach, assess, and reflect on a wide range of music content, skills, and knowledge, not just narrowly focused large ensemble performance. In doing so, the BEST program endorsed a vision of school music that was broad-based, inclusive, and focused on providing students with the skills to become lifelong learners in music.

Looking outside of our borders, we find that those nations with high-performing schools appear to be far less concerned with designing teacher evaluation systems and far more concerned with finding ways to support teachers in their work with students. In Finland, often pointed to as a model by leaders in the corporate reform movement, the national ministry of education is not involved in teacher evaluation, leaving this task to the teachers unions and local school governance bodies (Walker 2013b). Furthermore, Finnish teachers are not evaluated based on students' scores on standardized tests, because there are no standardized tests in Finland (Sahlberg 2010, 2).

There are several reasons for Finland's refusal to use standardized tests in its educational system. First, Finnish schools prize individualization and creativity over statistical indicators of student learning, Second, school leaders believe that the curriculum should drive instruction, not testing. Third, student assessment is seen as a school-level responsibility and not something that should be measured by externally administered and graded tests. Finally, "Finns believe that the problems often associated with external standardized testing—narrowing of the curriculum,

teaching to the test, and unhealthy competition among schools" are simply not worth the perceived benefits of testing (Sahlberg 2010, 7).

The consensus among the highest-performing schools internationally is that for teacher evaluation to be truly effective, it must emphasize "high-quality professional development, good working conditions, support from administrators, and a prominent role for teachers in designing new policies" (Walker 2013b). Conspicuously absent from this list is any mention of the use of student test scores to determine teacher effectiveness, or the use of teacher evaluation tools as a means to terminate "poor" teachers. Policymakers in these countries understand that when teachers are involved in designing teacher evaluation systems, we are more likely to see improvements in instruction. In the words of one teacher educator from Hong Kong: "Successful evaluation will help teachers think about students, and unsuccessful evaluation will make them think about themselves and their career" (Walker 2013b).

We Measure What We Treasure . . .

We often hear the old adage "We measure what we treasure" when discussions turn to issues of accountability in school curriculum offerings. This saying is typically used to provide justification for the narrowing of the curriculum to the disciplines of math and reading in the elementary grades, and to the science, technology, engineering, and math (STEM) subjects in the upper grades. Policymakers have focused on these subjects in part because the subject matter in these disciplines is conducive to measurement. According to Hargreaves and Braun, "Data driven improvement and accountability (DDIA) in the U.S. has focused on what is easily measured rather than on what is educationally valued. It holds schools and districts accountable for effective delivery of results, but without holding system leaders accountable for providing the resources and conditions that are necessary to secure those results" (2013, 24).

Hargreaves and Braun further point out, "Even if more and better data can be developed as a basis for improvement and intervention in the realms of education, health and social policy, there will still be limits to what DDIA[data-driven improvement and accountability] can accomplish" (2013, 5). Even the authors of a work entitled *Big Data* warn against the dangers of our obsession with data, stressing that there is a "special need to carve out a place for the human: to reserve space for intuition, common sense and serendipity" (Mayer-Schönberger and Cukier 2013, 196).

What could happen if our fascination with "big data" is left unchecked? In a stunning indictment of today's system of prioritizing accountability—and punishment—over improvement, Hargreaves and Braun offer the following summary of the outcome of current reform efforts:

> In the U.S., the high-stakes, high-pressure environment of educational accountability, in which arbitrary numerical targets are hierarchically imposed, has led to extensive gaming and continuing disruptions of the system, with unacceptable consequences for the learning and achievement of the most disadvantaged students. These perverse consequences include loss of learning time by repeatedly teaching to the test; narrowing of the curriculum to that which is easily tested; concentrating undue attention on "bubble" students near the threshold target of required achievement at the expense of high-needs students whose current performance falls further below the threshold; constant rotation of principals and teachers in and out of schools where students' lives already have high instability; and criminally culpable cheating. (2013, iii)

Rather than "treasuring what we measure," I would suggest that the opposite is actually true. For most of us, it is precisely those things that we value the most—our families, our friends, our students and colleagues, the beauty of a well-turned phrase, the joy of a student who has solved a musical problem—that are the most stubbornly resistant to being measured, and that the things we choose to measure are often chosen not for their value, but because they are easily measured. Tunks explains our quandary:

> Those of us involved in music education endeavors such as measurement, evaluation, and assessment would do well to heed the lesson of the inchworm. In our efforts to gain and retain acceptance in university and public school curricula we have adopted the "rules of the academy" and tried to show that the arts, specifically music, can be like other academic subjects. An important and justifiable aspect of this has been demonstrating that music education activities and programs can be evaluated successfully. After all, the message is clear that the subjects considered truly important in education are those that are evaluated. . . .
>
> Important risks, however, are associated with an inadequate perspective of measurement and evaluation in music. We could confuse the two, and fail in attempts to measure the unmeasurable. Or we could limit evaluation to an extension of measurement, and ignore the qualitative dimension that makes music unique and important in the first place. Still another possibility is that music educators, not having adequate evaluation tools, could simply abandon attempts at evaluation and lose the esteem of educational decision makers. (1987, 53–54)

It is our duty as educators and policymakers to be sure that the kinds of data we are using to evaluate teachers are not only valid and reliable, but meaningful and used appropriately. There is clearly great power in data, but there is also a responsibility to understand its limitations. We need to know when data is meaningful, and when it is not. As Hargreaves and Braun point out:

> This is equally true of those areas of life where problems are pervasive, inequities abound, and human suffering is rampant. Data can help in addressing these issues but in the end, some of our most challenging educational and social problems will not mainly be solved by more or better data, just as they will not be solved by more technology or by any other silver bullet. More and better data can help us make more efficient educational decisions and judgments, but they will not, of themselves, help us make wiser or more humane ones. Often, what we need to alleviate children's suffering and lack of opportunity is not more data or better metrics, but more attention, and more support. (2013, 6)

How can we apply this ethic to music education? The key would seem to be the combining of data with professional judgment. What is missing from much of the current dialogue in the reform movement is the voice of the teacher. Rather than using data to force music teachers into practices that do not align with their professional beliefs (e.g., using music to teach other subjects, assigning music teachers to supervise "credit recovery" classes in lieu of music classes, requiring all music teachers to develop classroom goals that support reading or math instruction), we should advocate for data use that encourages music teachers to apply their skills in new settings, and with new audiences.

For example, music teachers could work with school principals to create new courses and programs, such as guitar and songwriting classes, that would attract new students to school music offerings. Such a strategy would address one of the perennial shortcomings of most secondary school music programs, student participation rates that hover around 15 percent of school enrollment, while also representing a more culturally relevant approach to meeting students' interests and needs as music learners.

CONCLUSION

That we have a national education policy called Race to the Top should tell us everything we need to know about how wrong this approach truly is. Education is not a race. There should not be winners and losers in an educational system that is based on sound principles of equity, access, and fairness. Education is a process; it is a series of relationships carefully developed between teachers and learners and among learners. When education works well, everyone wins. That is not a race.

The phrase "to the Top" infers that every person is headed to the same place. However, education is not uniform—it is nuanced and individualized. Each learner has different goals, abilities, strengths, and weaknesses. Assuming that each student is headed to the same final destination on his or her educational journey is naive and uninformed.

The Race to the Top program has pitted each state against all the others, turning education into a game show style of competition. But education works best when we all work together, not against one another, and when we celebrate our differences instead of pretending that we are all the same, with identical goals and destinations. Education is not a race to the top. It is a journey of discovery—and finding out where each one of us is going is the goal. Music education can be a vital, critical component of each child's journey if we can work together to develop policies that support and encourage comprehensive musical experiences for all of our nation's students.

REFERENCES

Aguilar, C. E., and L. K. Richerme. 2014. What is everyone saying about teacher evaluation? Framing the intended and inadvertent causes and consequences of Race to the Top. *Arts Education Policy Review* 115 (4): 110–20.

American Education Research Association and National Academy of Education. 2011. Getting teacher evaluation right: A brief for policymakers. Accessed May 13, 2014, at https://edpolicy.stanford.edu/sites/default/files/publications/getting-teacher-evaluation-right-challenge-policy-makers.pdf.

American Statistical Association. 2014. ASA statement on using value-added models for educational assessment. Accessed May 13, 2014, at http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf.

Amrein-Beardsley, A. 2008. Methodological concerns about the education value-added assessment system. *Educational Researcher* 37 (2): 65–75.

———. 2014. *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York: Routledge.

Ballou, D. 2009. Test scaling and value-added measurement. *Education Finance and Policy* 4 (4): 351–83.

Berliner, D., and B. J. Biddle. 1995. *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley.

Bragdon, S. 2013. Teach For America founder Wendy Kopp: "Education is the tool to get kids out of poverty." *Generation Progress*, October 3. Accessed May 20, 2014, at http://genprogress.org/voices/2013/10/03/22243/teach-for-america-founder-wendy-kopp-education-is-the-tool-to-get-kids-out-of-poverty/.

Braun, H. 2005. *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Educational Testing Service.

Briggs, D., and B. Domingue. 2011. *Due diligence and the evaluation of teachers*. Boulder, CO: National Education Policy Center. Accessed October 9, 2012, at http://nepc.colorado.edu/publication/due-diligence.

Brustein, J. 2013. Microsoft kills its hated stack rankings. Does anyone do employee reviews right? *Bloomberg Businessweek*, November 13. Accessed May 20, 2014, at http://mobile.businessweek.com/articles/2013–11–13/microsoft-kills-its-hated-stack-rankings-dot-does-anyone-do-employee-reviews-right.

Duncan, A. 2009. Secretary Arne Duncan speaks at NEA conference; invites comments about test scores and teacher evaluations. *Homeroom* [blog], U.S. Department of Education, July 6. Accessed October 15, 2014, at http://www.ed.gov/blog/2009/07/secretary-arne-duncan-speaks-at-nea-conference-invites-comm/.

———.2011. School reform: A chance for bipartisan governing. *Washington Post*, January 2. Accessed May 20, 2014, at http://www.washingtonpost.com/wp-dyn/content/article/2011/01/02/AR2011010202378.html?referrer=emailarticle.

Halleck, M. 2013. What is stack ranking? Microsoft ends controversial employee rating system, Yahoo ramps it up. *International Business Times*, November 13. Accessed May 20, 2014, at http://www.ibtimes.com/what-stack-ranking-microsoft-ends-controversial-employee-rating-system-yahoo-ramps-it-1468850.

Hanushek, E., J. Kain, and S. Rivkin. 1998. Teachers, schools and academic achievement. Accessed May 9, 2014, at http://www.cgp.upenn.edu/pdf/Hanushek_NBER.PDF.

Hargreaves, A., and H. Braun. 2013. *Data-driven improvement and accountability*. Boulder, CO: National Education Policy Center. Accessed May 5, 2014, at http://nepc.colorado.edu/publication/data-driven-improvement-accountability/.

Hash, P. M. 2012. An analysis of high school band contest ratings. *Journal of Research in Music Education* 60:81–100.

———.2013. Large-group contest ratings and music teacher evaluation: Issues and recommendations. *Arts Education Policy Review* 114:163–69.

Hout, M., and S. W. Elliott, eds. 2011. *Incentives and test-based accountability in education.* Washington, DC: National Academies Press.

Kennedy, M. M. 2010. Attribution error and the quest for teacher quality. *Educational Researcher* 39 (8): 591–98.

Khadaroo, S. T. 2010. All teachers fired at R.I. school: Will that happen elsewhere? *Christian Science Monitor*, February 25. Accessed May 21, 2014, at http://www.csmonitor.com/USA/Education/2010/0225/All-teachers-fired-at-R.I.-school.-Will-that-happen-elsewhere.

Koedel, C., and J. Betts. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy* 5 (1): 54–81.

Mark, M. 2008. *A concise history of American music education*. Lanham, MD: Rowman & Littlefield Education.

Mathis, W. 2012. *Research-based options for education policymaking—section 1: Teacher evaluation*. Boulder, CO: National Education Policy Center. Accessed February 27, 2013, at http://nepc.colorado.edu/publication/options.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big data: A revolution that will transform how we live, work and think.* Boston: Houghton Mifflin Harcourt.

Michigan Arts Education Instruction and Assessment Project (MAEIA). 2013. MAEIA assessment specifications document. Accessed May 15, 2014, at http://mi-arts.wikispaces.com/file/view/ASD_-_Version_7_5_-_TOC-_Appendix_1_-_2.14.14%282%29.pdf/489679330/ASD_-_Version_7_5_-_TOC-_Appendix_1_-_2.14.14%282%29.pdf.

Michigan Society for Music Teacher Education. 2011. Position statement: Music teacher evaluation: Clarification and recommendations. Accessed May 13, 2014, at http://smte.us/wp-content/uploads/2011/11/mismte_position_-_statement_teacher_eval_-_final_draft.pdf.

Milner, H. R. 2013. *Policy reforms and de-professionalization of teaching*. Boulder, CO: National Education Policy Center. Accessed May 7, 2014, at http://nepc.colorado.edu/publication/policy-reforms-deprofessionalization.

National Association for Music Education (NAfME). 2011. Teacher evaluation: Position statement. Accessed May 9, 2014, at http://musiced.nafme.org/about/position-statements/teacher-evaluation.

———.2012a. NAfME recommendations for music teacher evaluation. Accessed May 9, 2014, at http://advocacy.nafme.org/files/2012/04/teacher_evaluation_ad_guide.pdf.

———.2012b. Teacher evaluation and assessment in an era of education reform [webinar]. Accessed May 9, 2014, at http://advocacy.nafme.org/webinar/teacher-evaluation.

———. 2013. New NAfME workbooks will aid in teacher evaluations. Accessed May 15, 2014, at http://musiced.nafme.org/news/new-nafme-workbooks-will-aid-in-teacher-evaluations/.

National Council on Teacher Quality. 2013. State of the states 2013: Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice. Accessed May 15, 2014, at http://www.nctq.org/dmsStage/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report.

Otterman, S. 2011. Closing schools have most challenging demographics. *City Room* [blog], *New York Times*, January 26. Accessed October 15, 2014, at http://cityroom.blogs.nytimes.com/2011/01/26/schools-chosen-to-close-have-toughest-demographics/?_php=true&_type=blogs&_r=0.

Phillips, A. M. 2011. 12 New York schools with low test scores are put on closing list. *New York Times*, December 9. Accessed May 20, 2014, at http://www.nytimes.com/2011/12/09/nyregion/12-new-york-schools-with-low-test-scores-are-put-on-closing-list.html.

Pink, D. H. 2009. *Drive: The surprising truth about what motivates us.* New York: Riverhead Books.

Polikoff, M. S., and A. C. Porter. 2014. Instructional alignment as a measure of teaching quality. *Education Evaluation and Policy Analysis.* Accessed May 13, 2014, at http://epa.sagepub.com/content/early/2014/04/11/0162373714531851.full.pdf+html?ijkey=Uwvo4Eg6.hQHI&keytype=ref&siteid=spepa.

Robinson, M. 1995. Alternative assessment techniques for teachers. *Music Educators Journal* 81 (5): 28–34.

Rothstein, R., R. Jacobsen, and T. Wilder. 2008. *Grading education: Getting accountability right.* Washington, DC: Economic Policy Institute.

Sahlberg, P. 2010. *The secret to Finland's success: Educating teachers*. Stanford, CA: Stanford Center for Opportunity Policy in Education. Accessed May 21, 2014, at https://edpolicy.stanford.edu/sites/default/files/publications/secret-finland%E2%80%99s-success-educating-teachers.pdf.

Sass, T. R. 2008. *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Policy Brief 4. Washington, DC: Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

Shaw, R. D. 2014. Music teacher evaluation in Michigan: Navigating choppy waters. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA, April 6.

Siebert. J. J. 2013. Policies and procedures in teacher assessment and evaluation. Accessed May 15, 2014, at http://musiced.nafme.org/files/2013/07/Teacher-Eval-NAfME.pdf.

Springer, M. G. 2009. *Performance incentives: Their growing impact on American K-12 education.* Washington, DC: Brookings Institution Press.

Tennessee Parents. 2014. Kids punished for poor test scores. February 2. Accessed May 20, 2014, at http://www.tnparents.com/2/post/2014/02/kids-punished-for-poor-test-scores.html.

Thomas, P. 2012. *Systemic poverty, the psychology of poverty, and misleading binaries*. Boulder, CO: National Education Policy Center. Accessed May 20, 2014, at http://nepc.colorado.edu/blog/systemic-poverty-psychology-poverty-and-misleading-binaries.

Tunks, T. W. 1987. Evaluation in music education: The value of measurement/the measurement of value. *Bulletin of the Council for Research in Music Education* 90:53–59.

Walker, T. 2013a. What do the 2012 PISA scores tell us about U.S. schools? *NEA Today*, December 3. Accessed May 20, 2014, at http://neatoday.org/2013/12/03/what-do-the-2012-pisa-scores-tell-us-about-u-s-schools/.

———.2013b. How do high-performing nations evaluate teachers? *NEA Today*, March 25. Accessed May 20, 2014, at http://neatoday.org/2013/03/25/how-do-high-performing-nations-evaluate-teachers/.

Warren, T. 2013. Microsoft axes its controversial employee-ranking system. *Verge*, November 12. Accessed May 20, 2014, at http://www.theverge.com/2013/11/12/5094864/microsoft-kills-stack-ranking-internal-structure.