Race to the Top FLDOE Developed Student Growth Models for Hard-to-Measure Course Content Areas (Music, Visual Arts, and Physical Education): Phase II

Volume I: Recommendations and Implementation

Florida Department of Education
Divison of Accountability, Research and Measurement
Office of Race to the Top Assessments
Race to the Top FLDOE Developed Student Growth Models for Hard-to-Measure Course Content Areas (Music, Visual Arts, and Physical Education): Phase II

Final Report

Prepared for:
Florida Department of Education
Division of Accountability, Research and Measurement
Office of Race to the Top Assessments

Project Number: 171-RG111-4C

TAPS Number: 14AT47

May 31, 2015

Prepared by:
Melanie DiLoreto, Ph.D., Kristen Wheatley, Valerie J. Morganson, Ph.D., Joanna Assefa, Sadie O’Neill, Daryll McKinley, Giang-Nguyen Nguyen, Ph.D., Steven J. Kass, Ph.D.*

Acknowledgements:
Julie Gray, Danny Drost, Rick Glaze, Sandra Durr, Reynaldo Valdez, Meiga Loho Noya, Mylene Biddle

University of West Florida, Center for Applied Psychology
11000 University Parkway, Pensacola, FL 32514

*Contact corresponding author at skass@uwf.edu or (850) 474-2107
Executive Summary

This report is a final summary of the activities of the second phase of a two-phase project concerning best practices for assessing student performance in the hard-to-measure courses of music, visual arts, and physical education in kindergarten through twelfth-grade (K-12). The two-phase project has been a partnership between the Florida Department of Education (FLDOE) and the University of West Florida (UWF). The first phase of the grant was approximately six months in duration (mid-December 2013 to June 2014) and focused on identifying the assessment landscape in the United States by (1) gathering quantitative and qualitative data concerning assessment practices in each of the 50 states and the District of Columbia and (2) reviewing extant literature on the various types of assessment methods and models. The results suggested a portfolio approach to assessment and greater focus on standards-based assessment. The second phase has spanned approximately eleven months in duration (August 2014 to June 2015) and has focused on four specific aims, outlined below (see Appendix A for timeline). Volume I of this report focuses on literature and research-based recommendations, gleaned from the literature and research summarized in Volume II. As related supplements, we also provide three “Considerations for Implementation” pamphlets – one for each hard-to-measure area – that can be utilized to communicate information to the various Florida districts. Volume II is the largest component of the report and is primarily for reference. It is chronologically organized based on the aims below and final recommendations.

1. **Aim One.** The first aim of the proposed research was to identify student learning outcomes and/or appropriate grade level expectations that are based on extant standards. We sought standards that would have face and content validity (i.e., they are perceived as having “clout” within the community of stakeholders, including teachers, administrators, and policymakers). We sought to adapt the standards as necessary (vs. adopt) to suit the needs of Florida’s various districts. Toward this end, as described in Chapter One, we carefully juxtaposed the Florida Next Generation Sunshine Standards and major national standards from Phase I of the grant and surveyed Florida stakeholders from the various districts to distill critical content for assessment.

2. **Aim Two.** The second aim of the current proposal was to provide a third-party review of the State’s pilot data (provided to UWF by the FLDOE) of the objective measures of hard-to-measure content areas. As described in Chapter Two, we reviewed psychometric properties of the data and provided item-level review of test content.

3. **Aim Three.** The third aim was to identify and recommend to the FLDOE various performance measures that can be used to assess music, visual arts, and physical education. This aim builds most directly on Phase I of the grant as it includes literature review of specific assessment practices. In addition to literature review, Aim Three included qualitative data collection to incorporate perspectives from Florida stakeholders. Chapter Three and Four summarize the results of Aim Three.

4. **Aim Four.** The fourth aim is a recommendation of a comprehensive model of student assessment. As agreed, the final recommendation is contingent upon the soundness
of the objective data collected and considers practical constraints, recognizing that extensive validity studies may not be feasible. The final recommended model is informed by the culmination of literature-based best practice and data gathered from both phases of the project. The final model incorporates objective, performance-based, and affective-based aspects of student achievement. Chapter Five summarizes how to compile measures and execute portfolio assessment according to the literature. Finally, Chapter Six discusses model components and proportions for each hard-to-measure area.

**Parties and Personnel**

It is important to note that this project is part of a larger series of FLDOE grants. The American Institute of Research provided assessment recommendations, which we reviewed as context for Phase II with the understanding that the final model must include a student growth component. In addition, there was a prior Race to the Top Performing Arts Assessment grant, led by Mary Grace Gordon and her team (2011 to 2014). We sought to build upon the viable parts of this prior work. Our own grant team was cross-disciplinary in nature. Co-principal investigators collectively have expertise in assessment and project management related to both education and workforce development. Our team included at least one subject matter expert from each hard-to-measure area, most of whom had worked with the State of Florida on prior projects related to K-12 assessment practices within their discipline. In addition, graduate students in the fields of Education, Industrial/Organizational Psychology, and Exercise Science have provided essential support throughout the project. Throughout the report, we refer to parties involved in the project as we relied upon them for expertise, particularly in establishing content validity.
Volume I

Final Recommendations & Considerations for Implementation: Curriculum Guidelines & Comprehensive Model of Assessment
A flexible, competency-based approach is recommended to accommodate various implementation complexities.

Based upon the literature reviewed and summarized in Volume II of this report, this volume focuses on the practical implications for adopting this model. Within this volume, we provide general recommendations. The information provided in this document is intentionally flexible and broad in order to accommodate the diverse scenarios and complexities surrounding educational practices in hard-to-measure disciplines. Assessment practices might ideally be tied to particular courses. However, curricula in hard-to-measure areas are highly variable. For example, in some cases, students complete a course in keyboarding to meet the requirements for the arts. As another example, while some students complete general physical education classes that cover a broad curriculum, some students may take a class in weight lifting. Drawing from a brief search of current FL enrollment data, our research team uncovered 167 course names listed for visual arts, 306 for performing arts, and 166 for physical education. Even when courses share similar learning outcomes, they may be administered at different “dosage” levels. For example, some classes meet once per week and others meet three times per week. Beyond variability in the curriculum, there is a wide range of student ability that impacts assessment. For example, some students have taken music courses much of their lives while others are minimally exposed throughout their schooling. Likewise, students enter a physical education course at highly variable levels of fitness. The model and procedure suggested below is amenable to adaptation to give flexibility to accommodate various applications. The suggested process focuses on assessment of student knowledge, performance, and affect.
Although it was not the main focus of this project, a major contribution of the research summarized in Volume II was to identify content to be assessed in each of the three hard-to-measure areas. Specifically, to guide the assessment processes, our research team (herein referred to as “we”) developed overarching constructs from the existing national and state standards. In addition to identifying overarching competencies, we identified important skills under each overarching construct to represent the various hard-to-measure areas at various grade levels. The overarching constructs and skills can be used for assessment. We further developed sample outcomes that can be measured. These are samples derived from existing standards and corresponding benchmarks and grade level expectations and can be further developed and refined and calibrated to be used in assessment.

In order to make an assessment recommendation to the FLDOE for the three hard-to-measure subject areas (i.e., music, visual arts, and physical education), we summarized information obtained from five sources: relevant literature, national survey results concerning current assessment practices, survey results of Florida professional educators, Florida subject matter expert opinions, and a thorough comparison of national and state standards. To address the two major foci of the grant, (1) recommendations for assessing performance and (2) development of a model of assessment, four Aims were constructed. Recommendations for each of the four Aims were garnered from a synthesis of the aforementioned sources. In the sections of this document that follow, a summary of the recommendations for each of the four Aims is provided.
Aim 1: Recommendations for the Outcomes that Guide Curriculum and Assessment

Although a major contribution of the present project was to synthesize national and state standards to extract commonality (see Aim One of Volume II), developing standards was not the main focus of this grant. Many of the subject matter experts that contributed to this project were participants in the initial drafts of the Florida Next Generation Sunshine State Standards (NGSSS). According to these subject matter experts, the purpose of the NGSSS was to guide curriculum – not assessment. As such, the standards and corresponding benchmarks are not written in a way that can be efficiently assessed and should be revisited. Furthermore, the standards should reflect measurable student learning or their comprehension of a given skill as a student progresses (e.g., as in Bloom’s Taxonomy) – an issue wherein the NGSSS sometimes fall short. In addition to developing overarching competencies and skills within each competency, we provide numerous example learning outcomes that may be considered by the FLDOE. As creating new standards and corresponding outcomes was not the original intent of the current project, the researchers drafted only samples that may be used as examples. However, these sample standards and outcomes are not to be construed as ready for implementation or to be published. Details of this process and the sample competencies and indicators are provided below in the Comprehensive Model of Assessment section for each of the respective disciplines (music, visual arts, and physical education).

The selection of national standards to be merged with the Next Generation Sunshine State Standards was a multi-step process that began in Phase I of the current project. Working from the perspective that other states’ experiences may be able to provide some insight into the process, preliminary searches of their Department of Education websites were conducted. If
national standards were used as a framework to develop their own standards, this was noted and then confirmed through direct interviews with state-level administrators. At the onset of Phase II, the most commonly cited national organizations that were favorably mentioned in the interviews were identified and their standards were thoroughly discussed by our subject matter experts. Each subject matter expert examined the pertinent standards in order to find those that created a best fit between Florida’s and the national association’s standards. The standards discussed were taken from the following: National Association for Music Education (NAfME), National Coalition for Core Arts Standards (NCCAS), and National Association for Sport and Physical Education (NASPE). The standards of these associations were determined to have the most skill alignment despite some organizational differences.

**Recommendation: Create an assessable set of standards or competencies for Music, Visual Arts, and Physical Education.**

In order to accomplish this, it is recommended that the FLDOE:

1. create a team of assessment, curriculum, and subject matter experts; and
2. facilitate the process necessary to create a comprehensive list of competencies and skills and/or learning outcomes that merge the existing state and national standards. When these goals are accomplished, these outcomes can then be used to effectively guide the assessment tasks, activities, and processes.

**Aim 2: Recommendations for Assessment Measures to use to Assess Student Knowledge (Cognitive Test Item Development)**

Our research team was given pilot data and access to the FLDOE Item Bank and Testing Platform (IBTP) to provide a third party review of data as they relate to one of the goals of the
grant: assessing student knowledge of music, visual arts, and physical education. As further described in Volume II, Aim Two, we used a sample of the data in the database to develop an overview of each test’s quality. In order to develop a meaningful overview, we focused on tests that covered a general area of the content. In light of the small amount of useable test data, limited demographic diversity of the sample, and the quality of the tests in relation to test development standards, we recommend that the testing database be carefully evaluated and re-piloted before it is used to make educational decisions.

There are multiple recommendations for consideration by the FLDOE related to testing the knowledge of students in the hard-to-measure disciplines.

**Recommendation for analyzing test data:** The FLDOE should require that each district develop the capability of statistically analyzing data obtained from these standardized tests. This can be accomplished by purchasing image scanner test scoring machines (e.g., Scantron) and by having a data analyst/statistician on staff, and/or sending out student responses for external analyses.
Aim 3: Recommendations for Specific Measures used to Assess Student Performance (and Affect)

Both national and state surveys were completed in order to collect information about current assessment practices. Furthermore, an in-depth review of best practices and empirical evidence found in the literature was conducted before final recommendations were developed. Finally, a group of subject matter experts was convened in order to provide additional evidence of various measures currently used to assess student performance. Multiple measures and methods of assessment were provided via these sources. Some examples include recorded performance assessment (audio, video, etc.), student logs, student projects, student journals, sketchbooks, fitness logs, fitness plans, and direct observations.
Sample artifacts include:

- **Checklists & rubrics** – used as tools by multiple raters to evaluate specified criteria
- **Student projects/work samples** – used as artifacts to demonstrate knowledge and skills in the discipline
- **Personal fitness plans** – used as a way for active student involvement in nutrition, health, and lifetime fitness; students monitor their progress toward goals that they have created with help from their teacher
- **Performance records/event tasks** – scores or other evidence to demonstrate appropriate skill and technique acquisition; performed during a single instructional period
- **Student logs** – used as a record of participation in various activities that demonstrate actions over time; also used as a means to document progress toward goals
- **Student journals** – used as a record of students’ reflections about their performance, attitudes, and feelings toward the discipline
- **Student responses to items on questionnaires** – used as artifacts to demonstrate appreciation and value judgement of the discipline

**Recommendation:** In order to demonstrate progress over time, a variety of at least eight to twelve artifacts should be collected at multiple points throughout the academic year and compiled into a process-based portfolio. For example, it is important that baseline evidence is captured and documented in a portfolio during the early days of the school year. Then, at a minimum, various artifacts should be collected at least two additional times during the year. By collecting multiple pieces of evidence at various points throughout the year, it is more likely that student progress can be effectively evaluated.
Aim 4: Recommendations for Implementing a Comprehensive Model of Assessment

Multiple assessments over time should be used in the evaluation of a student’s entire capacity to become a literate individual. Furthermore, authentic assessments that address students’ attitudes and performances that demonstrate literacy in the discipline are imperative to a successful evaluation of the teaching and learning process. As such, portfolio assessment provides a means by which student learning, skills, and attitudes can be evaluated by multiple raters. Although a portfolio itself is not an assessment, it does provide a representative collection of evidence that can be used to assess the acquisition of knowledge and skills that have occurred over time. It is important to note that the portfolio may include cognitive assessments regardless of whether or not the DOE implements a standardized test in each of the disciplines. If a standardized test is implemented across the state, then the evaluation of knowledge will be comprised of both the standardized test results as well as the demonstration of knowledge within
the portfolio. The combination of scores achieved from the standardized test and the portfolio assessments will comprise the overall percentage appropriate for the cognitive domain.

**Recommendation:** A comprehensive assessment of the hard-to-measure disciplines should encompass evidence from three domains (cognitive, affective, and performance).

**Three Domains of the Assessment Model**

**Domain 1: Cognitive Assessment** – Knowledge-based. Students complete a written test in order to demonstrate their knowledge about various concepts related to the discipline.

**Domain 2: Affective Assessment** – Attitudinal/Value-based. Students complete reflections and self-assessments regarding their feelings about and experiences in the development of social and personal behaviors. These assessments reflect students’ ability to work responsibly in both individual and group settings.

**Domain 3: Performance Assessment** – Skills & Techniques. Students apply skills and techniques in various activities necessary to demonstrate literacy in the discipline.

**Model Weighting**

Based on a comprehensive analysis of the recommendations found in the literature, national organizations, state and national benchmarks and grade level expectations, previous consultants, and our team’s subject matter experts, we recommend that the performance domain carry the most weight in the hard-to-measure disciplines. Although the performance domain should be the primary emphasis of assessment in each of the three disciplines, both the cognitive...
and affective domains are equally important in order to assess the whole child. Below, we explain the recommended weights for each of the domains for each of the three disciplines.

*Music and Art Education.* Music and art curricula emphasize the importance of performing, creating, and responding to the arts. Based on the consensus of multiple stakeholders regarding these aspects, our team recommends that the weighted breakdown of the model for comprehensive assessment of the arts is:

1) Performance – 40%
2) Cognitive – 35%
3) Affective – 25%

*Physical Education.* Physical education curriculum emphasizes the importance of demonstrating identified skills and techniques in addition to knowledge and attitudes about the subject matter. As such, our team recommends that the weighted breakdown of the model for comprehensive assessment of physical education is:

1) Performance – 45%
2) Cognitive – 35%
3) Affective – 20%
Evaluation Methods and Personnel

There are multiple ways to evaluate student learning and skill acquisition. As indicated above, it is important to include several pieces of evidence over time in order to document student progress. Furthermore, it is important to use evaluations from a variety of raters. Below is a list of possible means that can be used to conduct performance-based assessments.

Comprehensive Assessment of Student Learning

Recommendation: To ensure that each model component is represented, there are three options:

(Option 1) If objective testing is continued (see recommendations under Aim 2 above), test results should align with standards and may be used to represent the cognitive model component. Thus, portfolio artifacts should complement student test data by focusing on affective and performance model components and representing different learning outcomes.

(Option 2) If objective testing is discontinued, portfolio artifacts should represent all three model components and important learning outcomes based upon standards.

(Option 3) If objective testing is continued, cognitive portfolio artifacts may be used in combination with the objective test results for a sum of the overall cognitive weight of the model.
**Peer evaluations** – using checklists and rubrics, peers evaluate various criteria related to students’ demonstrated skills and techniques in various activities in order to enhance teaching and learning – also used as a means by which students demonstrate social responsibility in evaluating peers.

**Self-assessment** – using multiple formats (journals, logs, checklists, rubrics, etc.) students evaluate their own knowledge, skills, performance, and attitude toward the discipline.

**Teacher observation** – using checklists and rubrics, teachers evaluate student knowledge, skills, techniques, performance, and appreciation for the discipline.

**Expert evaluation and/or observation** – using checklists and rubrics, outside experts evaluate student knowledge, skills, techniques, performance, and appreciation for the discipline. These experts can be a cadre of master teachers, school, or district administrators. Technology may be utilized as a way of observing individual and/or class-wide demonstration of specified criteria (live feed; video; raters for electronic or hard-copy portfolio or a random sample of artifacts).
Scoring

The means by which scores are derived are generally tied to the overall purpose of the assessment. There are several considerations which should be taken into account when deciding the best way to score a process-based portfolio (i.e., a portfolio containing artifacts that reflect the evolution of student performance over time). A detailed explanation of scoring options is provided in Volume II.

Recommendation for Evaluating Student Artifacts: Because teachers will be most familiar with the learning outcomes in their course and the context of their classroom, they provide an important rating source. However, if student performance data are to be used to reflect teacher performance, relying upon teacher ratings may become problematic. In light of this, we suggest two options:

(Option 1) A combination of each of the above rating sources is included in the portfolio. Self-assessments can be used to assess the affective elements of the portfolio. Peer assessments can be used to assess soft skills such as teamwork where relevant. Teacher ratings will represent an important rating source, but are subject to random rating audits from experts. To prevent inflation of student performance ratings, accuracy and/or discrimination in ratings can be rewarded as an aspect of teacher performance.

(Option 2) The teacher works with a district representative to share course plans and provide a list of example assignments to be completed at the start of the school year. The district representative will sample artifacts to assess student performance. The teacher is unaware of which artifacts will be pulled for student evaluation.

In both Options 1 and 2, artifacts must be selected to represent important learning outcomes and model components (cognitive, performance, and affective).
Sample Rubrics & Checklists for Evaluating Individual Performance

Rubrics and checklists are popular and effective tools that can be used as a way to evaluate student performance. Rubrics describe specific criteria of levels of performance to assist in judging the quality of performance on a given task. There are two types of rubrics: holistic (i.e., providing a single rating of student performance) and analytic (i.e., providing a breakdown of student performance across multiple performance dimensions). Checklists are an efficient way to document skills at a given point in time. They are useful for tracking performance over time. Various sample rubrics and checklists are provided in the Appendices.

Recommendation: The FLDOE should determine the best way to score the process-based portfolio based on the stakes associated with the portfolio. If using the portfolio to make comparisons, scaled-scores may be the best approach. If using the portfolio to make decisions about student learning, then raw and/or percentages should be used.
Considerations for Implementing the Portfolio

Process portfolios are an effective way to capture student progress over time. Not only are many teachers aware of implementing portfolios because of the good cause exemption in the state of Florida, many have implemented portfolios in their own classrooms. As identified in Phase I of this grant project, our survey of practices in the United States and review of literature supports the implementation of portfolios as a common and best practice. Furthermore, portfolio assessment offers flexibility to educators and to students. Finally, portfolio assessment includes a variety of artifacts that allow students to demonstrate in multiple ways what they know and can do. Below are some suggested processes that the FLDOE may adopt as part of the comprehensive assessment of student knowledge, performance, and affect in these hard-to-measure disciplines.

Obviously there are several acceptable means by which to assess student learning and performance. As such, we have outlined a process that we believe is the best way to accomplish the goal of assessing student performance. Of course, there are pros and cons to using the processes described below that concern reliability of scores, training, resources needed, and logistics. Specifically, reliability of scores is often a concern anytime the activity that one is trying to assess is subjective in nature. In this case, expected performance can be construed as subjective. In order to combat this perceived subjectivity, it is important that multiple raters be used when possible. There are some disadvantages to using multiple raters, however. A
significant number of resources (both personnel and financial) are necessary. Professional development training is required in order to train all of the raters. Finally, logistics can be a problem when trying to deploy multiple raters to multiple sites.

Technology is one way to combat some of the logistics concerns. There are multiple uses of technology in performance assessment. For example, students may upload their musical performances (video and/or audio) to a purchased software system that is accessible to multiple raters. This process allows the raters to evaluate performance at their convenience – not necessarily at the time of the live performance. Another way that technology can be used is by the teachers and district representatives to complete rubrics electronically. The use of this sort of technology allows for quick analyses because the performance results are immediately captured electronically. Finally, videos can be used in the classroom for later review by the teacher and/or district administrators. If a class of students is video-taped while they are all performing a skill or technique, then multiple raters can watch the video at a later time in order to evaluate each student’s performance.

District leaders and classroom teachers should decide the types of artifacts that can be used to effectively demonstrate knowledge, performance, and affect. This will allow for the adaptation of the portfolio to meet a specific circumstance. For example, a physical educator who teaches several classes ranging from general fitness to weightlifting should provide input to the specific artifacts that are required for the portfolio. Some general recommendations include having a minimum of eight artifacts that are taken from student projects with corresponding rubrics completed by multiple raters, student journals, and student reflections. The artifacts themselves then will allow for assessment of each of the three domains (cognitive, performance,
and affect). For example, a student completes a project that demonstrates his performance using a specified technique. The teacher and a district administrator observe the student’s performance and each rater completes a pre-established rubric. The student then evaluates his own performance and writes responses to pre-determined reflective questions. The student evaluation and reflection is captured in his journal. Then, the portfolio contents for this one performance task become: (1) the student project (evidence of performance); (2) the teacher completed rubric and the district administrator completed rubric (evidence of performance); and (3) the student journal that includes the self-evaluation and the reflection (evidence of affect). This same process is used throughout the entire academic year in order to demonstrate progress over time across these three domains. At the end of the year, the portfolio contents are evaluated by multiple evaluators (who are selected by the district). There are multiple ways to accomplish the goal of having multiple evaluators. First, the portfolio contents may be captured in a folder (electronic or physical). At the end of the year, the district can evaluate all contents in the portfolio for a random sample of students or the district may choose one or more artifacts to review based on the emphasis of learning for that particular school year. Either approach requires the use of a pre-established portfolio evaluation rubric that would be used by multiple raters in order to determine inter-rater reliability of the portfolio evaluation. For example, the classroom teacher evaluates each of his/her student’s portfolios using the pre-established process portfolio rubric. Then, one or more trained district evaluators also score the portfolio using the same rubric. The raters’ scores are then compared in order to establish reliability of the results.

Validity of the scores is also a concern when trying to assess student knowledge, performance, and affect. An important step that can be employed in order to increase the validity of the scores is to ensure a close alignment of the artifacts to the identified learning outcomes. It
is extremely important that the FLDOE or the districts specify the process as much as possible. Given the multiple options for courses and circumstances described above, this will pose a significant challenge. As such, it is recommended that specific outcomes for assessment be required for all students regardless of the circumstance. In primary school, curricula are more consistent because hard-to-measure subjects are less likely to be an elective and standards are more consistent and concrete (see Volume II for a full summary). These factors will aid the implementation of a portfolio system in early education. In middle and/or secondary school, students could be assigned the requirement of completing the portfolio to address the identified outcomes. In this case, the students would compile evidence and have teachers evaluate their performance and complete the rubrics (scheduled at agreed-upon times). In this approach, the responsibility is placed on the students to demonstrate the requirements instead of the teachers capturing the data from the students. This model is similar to requirements for home-schooled students. The responsibility to demonstrate knowledge and performance is placed on the students’ families instead of on a school. If every student is required to demonstrate competency on the identified outcomes regardless of the circumstances, the validity of the scores will likely improve. It should be noted that the less standardized the process becomes, the likelihood of obtaining valid results diminishes.