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 II: Literature Review and Methods

Florida Department of Education

Division 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
# Table of Contents

Chapter One: Aim One ................................................................................................................................... 5  
  Background Information ........................................................................................................................... 5  
  Methodology ............................................................................................................................................. 8  
    Procedure .............................................................................................................................................. 8  
    Participants ........................................................................................................................................... 9  
  Survey Results ........................................................................................................................................ 17  
    Music Education ................................................................................................................................. 18  
    Art Education ...................................................................................................................................... 20  
    Physical Education .............................................................................................................................. 23  
Chapter Two: Aim Two .................................................................................................................................. 27  
  Background Information ......................................................................................................................... 27  
  Analysis Results ...................................................................................................................................... 29  
    HtM_Personal Fitness_1501300_FTS14_Form#1 .............................................................................. 31  
    HtM_Hope Core_3026010_FTS14_Form#1 ...................................................................................... 32  
    HtM_Team Sports 1_1503350_FTS14_Form#1 ................................................................................ 33  
  Conclusions and Interpretation ............................................................................................................... 34  
Chapter Three: Aim Three – Input from FL Stakeholders .............................................................................. 37  
  Method .................................................................................................................................................... 37  
  Design ................................................................................................................................................. 37  
  Participants and Recruitment .............................................................................................................. 38  
  Procedure ............................................................................................................................................ 39  
  Results ..................................................................................................................................................... 40  
    Music Education ................................................................................................................................. 41  
    Visual Arts Education ......................................................................................................................... 41  
    Physical Education .............................................................................................................................. 42  
Chapter Four: Aim Three – Literature-Based Inquiry .................................................................................... 43  
  (1) Recorded Performance Assessments ................................................................................................. 43  
    Music Education ................................................................................................................................. 43  
    Visual Arts Education ......................................................................................................................... 44  
    Physical Education .............................................................................................................................. 44  
  (2) Student Logs ...................................................................................................................................... 50  
    Music and Visual Art Education ........................................................................................................ 51
Physical Education................................................................. 52
(3) Sketchbooks ................................................................. 53
(4) Fitness Plans ................................................................... 54
(5) Self-Assessment/Reflection ........................................... 57
  Reliability and Validity of Self-Assessment ..................... 58
(6) Fitness Tests ................................................................. 65
(7) Checklists ....................................................................... 66
(8) Rubrics ........................................................................... 68
  Types of Rubrics ............................................................. 68
  Quality Rubrics ............................................................... 69
  Adapting an Existing Rubric ............................................ 71
  Developing a Rubric ........................................................ 72
  Rubrics by Hard-to-Measure Area .................................... 74

Chapter Five: Compiling Measures into Portfolio Assessment ................................................. 80
  Portfolio Assessment and Associated Artifacts ............... 80
  Rating and Evaluation ..................................................... 84
  Presentation & Storage .................................................... 89
  Conclusion ................................................................. 89

Chapter Six: Aim Four Model Development ............................................................................. 91
  Models of Assessing Students in Music Education, Art Education, and Physical Education 91
  Three Domains of the Assessment Model ....................... 91
    I. Assessing Student Knowledge ................................. 92
    II. Assessing Student Performance ............................ 93
    III. Assessing Student Affect ....................................... 94
  Weight Recommendations Based on Models of Assessment 95
  Findings from National Organizations and from the Literature .................................................. 95
  Previous Florida Race to the Top Grant (RTTT) Project Findings ........................................... 96
  Review of Existing Standards and Corresponding Benchmarks/Grade Level Expectations ...... 98
  Growth versus Proficiency ................................................ 99
  Raw and Scaled Scores.................................................... 100
  Summary ................................................................. 102
Chapter One: Aim One

Background Information

As a result of accountability demands placed on professional educators, there is a need for standardized methods to assess performance-based, hard-to-measure disciplines such as music education, art education, and physical education. We selected an approach of combining national standards with the Florida Next Generation Sunshine State Standards (NGSSS) and associated benchmarks. In doing so, the curricula may include content and process standards that districts, schools, and classroom educators may use to assess student performance. The following competencies are identified from literature and from a survey of discipline-specific professional educators as the most important for the three aforementioned disciplines.

Music Education. The National Association for Music Education (NAfME) purported that agreement on what students should know/perform is a central component to an effective music program. NAfME (2014) recommended, and our research team concurs, that the following nine national standards be employed as a basis for music education and its assessment: (1) singing, alone and with others, (2) performing on instruments, alone and with others, (3) improvising melodies, variations, and accompaniments, (4) composing and arranging music, (5) reading and notating music, (6) listening to, analyzing, and describing music, (7) evaluating music performances, (8) understanding relationships between music and the other arts, and (9) understanding music in relation to history and culture.
These standards focus on students’ ability to create, perform, respond, and connect to music. Additionally, these standards emphasize the opportunity-to-learn by recommending that all students:

… receive substantive, sequential, standards-based music instruction from expert music educators throughout grades PreK-8 as part of their core curriculum; have opportunities to elect additional music classes, such as ensembles, beginning in the intermediate grades; and have opportunities and encouragement to pursue one or more music course sequences for four years at the high school level.


Furthermore, students should have opportunities to conceive and develop new ideas and work, perform and present their work, respond by evaluating how music conveys meaning, and connect their work with personal meaning and external context (National Core Arts Standards, 2014).

**Art Education.** The National Art Education Association (NAEA) recommended two areas of competence for students: (1) content standards emphasizing what students learn/perform in the arts, and (2) achievement standards focusing on the levels of achievement students demonstrate at the completion of various grade levels (The National Visual Arts Standards, 1994).

NAEA also recommended that students in the early grades should be taught the differences between art materials, techniques, and processes (The National Visual Arts Standards, 1994). Furthermore, NAEA recommended that students in the higher grades be able to communicate proficiently in at least one form of art, including the ability to define and solve
artistic problems with insight, reason, and technical proficiency (The National Visual Arts Standards, 1994). Like music education, the National Coalition for Core Arts Standards (2014) recommended that students have opportunities to create works of art and develop new artistic ideas, present, interpret and share their artistic work, respond by evaluating how works of art convey meaning, and connect their works of art with personal meaning and external context. Our research team supports this approach and recommends a tiered assessment strategy for art education at the various grade levels.

*Physical Education.* The National Association for Sport and Physical Education [NASPE] (2007) recommended that physical education teachers base their teaching on the national standards for K-12 physical education. The amount of instructional time spent on physical education is a necessary component to an effective physical education program. NASPE (2007) suggested a minimum of 150 minutes per week for elementary-aged students and 225 minutes per week for middle and high school students for all physical education programs. Additionally, physical education teachers must not only establish high levels of expectations related to psychomotor, cognitive, and affective domains, but must also view assessment as an integral part of teaching and learning (NASPE, 2007).

According to the American Alliance for Health, Physical Education, Recreation and Dance “Grade-Level outcomes for K-12 Physical Education” (2013), the goal of physical education is to “develop physically literate individuals who have the knowledge, skills and confidence to enjoy a lifetime of healthful physical activity” (p. 1). In this respect, the individual must; (1) have learned the skills necessary to participate in a variety of physical activities, (2) know the implications and the benefits of involvement in various types of physical activities, (3)
participate regularly in physical activity, (4) be physically fit, and (5) value physical activity and its contributions to a healthful lifestyle. Skills development contributes to this longitudinal process and provides a means of assessment of not only students, but teachers as well. Thus, we recommend that skills development be implemented as a critical component of physical education as well as an assessment tool.

Methodology

Procedure

We used a variety of methods to compile information for its recommendations. By first becoming experts in the literature, various professional organizations and their affiliated standards, we organized a summary and history of each set of standards for each of the three disciplines (For timeline, see Appendix 1A). A detailed report of the various standards may be found in Appendix 1B. We became familiar with both the national and the state standards associated with each of the three disciplines, and developed crosswalks to compare national and state standards and benchmarks. A detailed report of how these comparisons were done, a crosswalks summary, and the percentage of overlap between national and state standards for each of the three disciplines may be found in Appendix 1C. Finally, we compiled tables that incorporated all of the benchmarks for each of the standards. These benchmark tables were used to help our team identify the progression of knowledge and/or skill needed in each grade level.

In addition to compiling information from the above listed sources, we surveyed current professional educators in the state of Florida. This was done to identify standards, competencies, and benchmarks that represent important curriculum aspects in K-12 education in the areas of
physical education, art, and music. Additionally, the survey sought to identify the extent to which standards, competencies, and benchmarks are primarily knowledge-based versus performance-based in order to ensure the content validity of future research and assessment practice.

Letters of invitation were emailed to 65 of Florida’s 67 school districts plus Florida Virtual School, University of Florida Lab School, Florida School for the Deaf and the Blind, Florida State University Lab School, and Florida Agricultural and Mechanical School District. The emailed invitations were sent to district level physical education and fine and performing arts curriculum coordinators/supervisors serving as primary contacts. These primary contacts were asked to forward the letter of invitation to their physical education, visual arts, and performing arts teachers. Although the number of participants to whom the questionnaire was sent is unknown, representatives from 34 districts completed the questionnaire with highest representation from Miami-Dade (20.4%), Palm Beach (10.9%), Bay (9.3%), and Duval (9.1%) (see Appendix 1D).

Participants

Demographics. There were a total of 1005 participants that accessed the questionnaire, with 660 completing the questionnaire and whose data were used in the analyses. The breakdown of participants by grade level for music, art, and physical education is shown in Table 1-1. A total of 193 participants completed the survey indicating their expertise in music education. In the art education discipline, 289 participants provided complete data and were retained in the sample. A total of 178 participants provided complete data and were retained in
the sample. Figure 1-1 illustrates the number of participants in each disciplines and how each portion makes up a percentage of our total 660 participants.

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Music Education</th>
<th></th>
<th>Art Education</th>
<th></th>
<th>Physical Education</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency (n)</td>
<td>Percent</td>
<td>Frequency (n)</td>
<td>Percent</td>
<td>Frequency (n)</td>
</tr>
<tr>
<td>K</td>
<td>8</td>
<td>4.1</td>
<td>10</td>
<td>3.5</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>4.7</td>
<td>5</td>
<td>1.7</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>8.3</td>
<td>14</td>
<td>4.9</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>17.6</td>
<td>41</td>
<td>14.2</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>6.2</td>
<td>25</td>
<td>8.7</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>13.0</td>
<td>47</td>
<td>16.3</td>
<td>42</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>7.8</td>
<td>20</td>
<td>7.0</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>4.7</td>
<td>15</td>
<td>5.2</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>10.4</td>
<td>23</td>
<td>8.0</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>5.2</td>
<td>15</td>
<td>5.2</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>6.2</td>
<td>15</td>
<td>5.2</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>3.1</td>
<td>26</td>
<td>9.0</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>8.8</td>
<td>33</td>
<td>11.4</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>193</td>
<td>100</td>
<td>289</td>
<td>100</td>
<td>178</td>
</tr>
</tbody>
</table>
Figure 1-1: Pie chart showing the percentage breakdown of total participants by discipline.

Figure 1-2 illustrates the participants’ allocation of responses concerning the level of schooling in which they taught (represented in each discipline). Primary/elementary school was the most common level of schooling for participants in all disciplines. Figure 1-3 illustrates the number of schools taught at by participants in each discipline (i.e., one school, two schools, and three or more schools). The majority of music education participants taught at one school, while the majority of participants in art education and physical education taught at two schools.
Figure 1-2: Bar graph showing the level of schooling taught by participants in each discipline.

Figure 1-3: Bar graph showing number of schools taught at by participants in each discipline.

Figure 1-4 illustrates the type of community the participants in each discipline primarily taught in (i.e., urban community, suburban community, rural community, or low income community). As illustrated in Figure 1-5, the large majority of participants in each discipline indicated teaching in public schools rather than private and charter or magnet schools.
Figure 1-4: Bar graph showing type of community taught in by participants in each discipline.

Figure 1-5: Bar graph showing the type of school taught in by participants in each discipline.

Figure 1-6 depicts the average number of years of teaching experience of participants in each discipline. As shown, music and art educators tended to have more experience than physical educators. Figure 1-7 illustrates the highest degree obtained by participants in each discipline with the vast majority of participants in each discipline holding a bachelor’s or master’s degree.
**Figure 1-6:** Bar graph showing the average years of teaching experience of participants in each discipline.

![Average Years of Experience Teaching](image1)

**Figure 1-7:** Bar graph showing the highest degree obtained by participants in each discipline.

![Highest Degree of Education](image2)

**Figure 1-7:** Bar graph showing the highest degree obtained by participants in each discipline.

Figure 1-8 illustrates that a majority of participants in music, art, and physical education indicated that their degree is in the discipline in which they teach. A minority of participants in music education indicated that their degree is in a field closely related to music (e.g., minor in music, music and educational leadership, piano, vocal performance) and more participants in
indicated that their degree is in another field (e.g., art education, education, general K-6, history, theatre and dance, science). Likewise, a minority of participants in art education hold a degree in a closely related field (e.g., indicating a minor in art education, master’s of fine art, bachelor’s of fine art, graphic design, ceramics, studio art, interior design) or hold a degree in another field (e.g., elementary education, general K-6, information technology, theatre, English, business). In physical education, a minority of participants hold degrees in other fields (e.g., health science, business, elementary education, history, political science, mathematics, special education, social science) or hold degree in a closely related field (e.g., indicating a degree in exercise science, sports leadership, kinesiology, sports administration, recreation and leisure services).

Figure 1-8: Bar graph showing the field of degree held by participants in each discipline.

The majority of participants in each discipline received their teaching certification through traditional university teacher training programs (as shown in Figure 1-9). The next largest percentage of certification programs among the participants was a college or university alternative program. A minority of participants in all disciplines indicated receiving their teaching certification through a district alternative program.
Figure 1-9: Bar graph showing the type of teaching certification program from which participants in each discipline received their certification.

Figure 1-10 illustrates that the majority of participants in all disciplines hold professional certifications in the discipline in which they teach. Fewer participants in music education and physical education hold temporary certifications in the discipline in which the teach and the minority of participants in these disciplines hold certification in another area (e.g., drama, elementary education, professional music for music education) and (e.g., elementary education, clinical education, professional leadership education, professional Spanish). While most participants in the art education discipline hold professional certifications in the disciplines in which they teach, fewer hold certifications in other areas (e.g., drama, elementary education, business, computer science), and a minority hold temporary certifications in art education.
Figure 1-10: Bar graph showing the type of certificate held by participants in each discipline.

Survey Results

We grouped responses by school level: elementary (K to 5), middle (6 to 8), and high school (9 to 12). The decision to group the grade levels in this manner was based upon (a) our review of prior literature (e.g., Roslow Research Group, 2009), (b) a tendency for respondents to indicate that they teach across multiple grade levels within an educational level, and (c) our data analyses. Specifically, we ran multivariate analyses of variance (MANOVAs)\(^1\) for each of the three subject areas (music, art, and physical education) using grade level as a predictor and overall competencies as an outcome for the three major content areas and did not find a pattern of reliable differences by grade level. We also used MANOVA to evaluate whether elementary school should be split into two levels (K to 3 and 4 to 5), but the result was not statistically significant to conclude a reliable pattern of results to support doing so.

\(^1\) MANOVA is a statistical procedure for identifying reliable mean differences. It is utilized to test for reliable mean differences when a categorical variable (e.g., grade level) predicts a set of continuous variables (e.g., competency importance ratings).
Participants rated the overall importance of five general competencies on a scale ranging from zero to three with the following anchors: (0) not important, (1) somewhat important, (2) important, and (3) very important. Correspondingly, in the results reported below, we refer to competencies rated as 1.5 to 2.4 as “important” and those rated as 2.5 or higher as “very important.” They rated more specific topics nested within each of the overall competencies on two scales. The first scale for specific topics included four anchors: (1) not important, (2) somewhat important, (3) important and (4) very important. Correspondingly, we refer to specific topics rated 1.5 to 2.4 as “somewhat important,” those from 2.5 to 3.4 as “important,” and those 3.5 and above as “very important.” We also discuss standard deviations that were 1.0 or above prior to rounding as these may reflect rater unreliability. The second scale allowed participants to mark each specific topic as based on “knowledge” (i.e., it can be assessed using a written test) and/or “performance” (i.e., it would require some kind of demonstration to assess).

**Music Education**

*Competencies.* Based on combined national and NGSSS, the overarching competencies for music education in the survey were as follows: (1) thinking, reflecting, and responding to performing arts, (2) historical and global connections in performing arts, (3) creating innovative performing arts, (4) organizing and presenting performing arts, and (5) connecting skills, techniques, and processes in performing arts. Means and standard deviations are reported in Table 1-2. Across elementary, middle, and high school levels, the first four competencies can be characterized as “important” whereas the fifth competency was rated slightly higher and can be characterized as “very important.”
Table 1-2. Overall Competency Importance Ratings for Music Education

<table>
<thead>
<tr>
<th>Competency</th>
<th>Mean</th>
<th>SD</th>
<th>Rating Category</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elementary School (K to 5)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thinking, reflecting, and responding to music.</td>
<td>2.2</td>
<td>0.6</td>
<td>Important</td>
</tr>
<tr>
<td>Historical and global connections in music.</td>
<td>1.8</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Creating innovative music.</td>
<td>2.1</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Organizing and presenting music.</td>
<td>2.1</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Connecting skills, techniques, and processes in music.</td>
<td>2.7</td>
<td>0.5</td>
<td>Very Important</td>
</tr>
<tr>
<td><strong>Middle School (6 to 8)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thinking, reflecting, and responding to music.</td>
<td>2.1</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Historical and global connections in music.</td>
<td>1.6</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Creating innovative music.</td>
<td>1.7</td>
<td>0.8</td>
<td>Important</td>
</tr>
<tr>
<td>Organizing and presenting music.</td>
<td>2.0</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Connecting skills, techniques, and processes in music.</td>
<td>2.5</td>
<td>0.7</td>
<td>Very Important</td>
</tr>
<tr>
<td><strong>High School (9 to 12)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thinking, reflecting, and responding to music.</td>
<td>2.3</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Historical and global connections in music.</td>
<td>1.8</td>
<td>0.8</td>
<td>Important</td>
</tr>
<tr>
<td>Creating innovative music.</td>
<td>1.8</td>
<td>0.8</td>
<td>Important</td>
</tr>
<tr>
<td>Organizing and presenting music.</td>
<td>2.1</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Connecting skills, techniques, and processes in music.</td>
<td>2.5</td>
<td>0.6</td>
<td>Very Important</td>
</tr>
</tbody>
</table>

Specific Topics. Ratings for specific topics for elementary, middle, and high school levels can be found in the district pamphlets as well as Appendices 1E, 1F, and 1G, respectively. Across all schooling levels, each of the nine specific topics, on average, was rated as either “important” or “very important.” In order to identify which associated skills were most important to Florida educators, the mid-range of scores for each discipline was used to determine where to make the cut of importance. Across specific topics, even within one competency and schooling phase, participants endorsed a mixture of knowledge and performance-based assessment requirements.

Qualitative Data. Additionally, after answering structured questions, participants were asked to write in any additional competencies that they felt should be represented. A total of 37
individuals provided commentary. Our music education subject matter expert summarized the results of these write-in responses.

At the elementary school level, most of the comments pertained to reading music and learning rhythms. Five of the eleven comments addressed the lack of standards for learning to read music, both note names and rhythm. One participant felt that elementary students should focus on understanding and appreciating music rather than note reading, music history, and composers. Two respondents felt there should be standards for movement. There was a complaint about the standards being too broad and a comment that all teachers should be provided the same tools such as recordings, texts, and instruments. There was also a comment about relating music to the other arts and a comment about appreciating music.

The prevailing comments from middle and high school respondents dealt with assessing performance. Nine out of 13 middle school respondents mentioned performance and five out of 15 high school respondents mentioned performance. The respondents generally felt there was not enough emphasis on performance in the standards, although there seems to be a prevailing attitude that it is difficult to assess performance. There were some comments about relating music to other disciplines. Music etiquette was mentioned, but this is closely related to performance.

**Art Education**

**Competencies.** The overarching competencies for art education in the survey were as follows: (1) thinking, reflecting, and responding to visual arts, (2) historical and global connections in visual arts, (3) creating innovative visual arts, (4) organizing and presenting
visual arts, and (5) connecting skills, techniques, and processes in visual arts. Means and
standard deviations are reported in Table 1-3. Across elementary, middle, and high school
levels, “thinking, reflecting and responding,” “historical and global connections” and
“organizing and presenting” can be characterized as “important.” Two competencies, “creating
innovative visual art” and “connecting skills, techniques and processes” were rated slightly
higher and can be characterized as “very important.”

### Table 1-3. Overall Competency Importance Ratings for Art Education

<table>
<thead>
<tr>
<th>Competency</th>
<th>Mean</th>
<th>SD</th>
<th>Rating Category</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elementary School (K to 5)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thinking, reflecting, and responding to visual art.</td>
<td>2.1</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Historical and global connections in visual art.</td>
<td>2.0</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Creating innovative visual art.</td>
<td>2.6</td>
<td>0.5</td>
<td>Very Important</td>
</tr>
<tr>
<td>Organizing and presenting visual art.</td>
<td>2.1</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Connecting skills, techniques, and processes in visual art.</td>
<td>2.6</td>
<td>0.5</td>
<td>Very Important</td>
</tr>
<tr>
<td><strong>Middle School (6 to 8)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thinking, reflecting, and responding to visual art.</td>
<td>2.3</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Historical and global connections in visual art.</td>
<td>2.1</td>
<td>0.7</td>
<td>Important</td>
</tr>
<tr>
<td>Creating innovative visual art.</td>
<td>2.8</td>
<td>0.5</td>
<td>Very Important</td>
</tr>
<tr>
<td>Organizing and presenting visual art.</td>
<td>2.3</td>
<td>0.6</td>
<td>Important</td>
</tr>
<tr>
<td>Connecting skills, techniques, and processes in visual art.</td>
<td>2.7</td>
<td>0.5</td>
<td>Very Important</td>
</tr>
<tr>
<td><strong>High School (9 to 12)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thinking, reflecting, and responding to visual art.</td>
<td>2.4</td>
<td>0.6</td>
<td>Important</td>
</tr>
<tr>
<td>Historical and global connections in visual art.</td>
<td>1.9</td>
<td>0.8</td>
<td>Important</td>
</tr>
<tr>
<td>Creating innovative visual art.</td>
<td>2.7</td>
<td>0.5</td>
<td>Very Important</td>
</tr>
<tr>
<td>Organizing and presenting visual art.</td>
<td>2.2</td>
<td>0.8</td>
<td>Important</td>
</tr>
<tr>
<td>Connecting skills, techniques, and processes in visual art.</td>
<td>2.7</td>
<td>0.5</td>
<td>Very Important</td>
</tr>
</tbody>
</table>

*Specific Topics.* Ratings for specific topics for elementary, middle, and high school
levels can be found in the district pamphlets and in Appendices 1H, 1I, and 1J, respectively.
Across all schooling levels, each of the specific topics, on average, was rated as either
“important” or “very important.” However, “Studying music to increase 21st century skills
needed in a global economy” at the middle school level was rated at the low end of the
“importance spectrum” and in both middle and high school had lower rater agreement than the other specific topics. Across specific topics, even within one competency and schooling phase, participants endorsed a mixture of knowledge and performance-based assessment requirements.

**Qualitative Data.** Additionally, after answering structured questions, participants were asked to write in any additional competencies that they felt should be represented. A total of 87 individuals provided commentary for analysis. Our art education subject matter expert summarized the results of these write-in responses. The one theme that appeared across the three grade levels was an emphasis on careers, including preparation for employment in the real world and the consumer market (elementary = 4/22, middle = 4/33, high = 3/32). Elementary and middle school teachers both commented on the understanding and application of the Elements and Principles of Art Design (elementary = 3/22, middle = 4/33). Both elementary and middle school teachers touted the need for creativity, but elementary also added risk-taking and problem-solving (elementary = 4/22) and middle school leaned toward creativity and expression through personal art (middle = 6/33). Elementary teachers wanted more emphasis for students to build skills through the practice of art (elementary = 3/22). The high school teachers wanted to see more appreciation of art (high = 6/32) and instruction in and opportunity for portfolio building (high = 3/32) in the assessment.

Also, across the grade levels, there were multiple comments that did not address any missing competencies but dealt with frustrations associated with the lack of time for teachers to teach art and students to practice art and the difficulty in successfully teaching art given the numbers of students in art classes. There were also concerns about non-authentic assessment
where testing involves just the knowledge of art as a subject and not the doing of art and the perception that the testing does not relate to the standards anyway.

Physical Education

Competencies. The overarching competencies for physical education in the survey were as follows: (1) movement and motor skills, (2) concepts, principles, strategies, and or tactics related to a variety of physical activities, (3) health-enhancing level of physical fitness, (4) personal and social behavior in physical activity settings, and (5) values related to self and others in physical activity settings. Means and standard deviations are reported in Table 1-4. Across elementary, middle, and high school levels, all competencies were rated as either “very important” or “important.”
Specific Topics. Ratings for specific topics for elementary, middle, and high school levels can be found in the district pamphlets and in Appendices 1K, 1L, and 1M, respectively. The vast majority of the specific topics representing the five major competencies were rated as either “important” or “very important” by participants. However, several exceptions should be noted. First, aquatics was rated as “somewhat important” and the standard deviation was relatively high, indicating that it was not reliably rated. At the elementary school level,
biomechanics and resistance training were rated as only “somewhat important” on average. Likewise, at the elementary school level, while “fitness assessment/program planning” rated slightly higher than “somewhat important,” the standard deviation was relatively high. Conversely, dance and gymnastics were rated as only “somewhat important” at the high school level. “Striking with long implements” and “striking with short implements” were rated at the low end of the spectrum within the “important” band and demonstrated relatively low reliability.

**Qualitative Data.** Additionally, after answering structured questions, participants were asked to write in any additional competencies that they felt should be represented. Our physical education subject matter expert summarized the results of these write-in responses. At the elementary level teachers appear to be unfamiliar with the state standards and benchmarks. Additionally, teachers complained about areas that are part of the physical education curriculum and stated that they should be removed. These teachers should instead take the time and initiative to learn the content in order to provide the well-rounded physical education curriculum content. Another complaint at elementary levels was the inappropriateness of cognition in physical education. This type of statement suggests that these teachers have not had the formal training for teaching cognition in physical education. Cognition does not have to be taught separately from physical activity but in conjunction with movement, this requires more time and energy during the planning and teaching in physical education.

At the middle school level emerging themes included: (1) access to non-competitive experiences, (2) opportunities to be creative during physical education, and (3) more advanced game and sport tactics and strategies. The middle school teachers’ utilized student-designed games, this is traditionally done in elementary school suggesting that teachers may be more interested in less competitive games. In regards to sport, this would call for more advanced drill
and practice planning that would also enable students to use their skills in sport outside of school and later in life.

The high school level themes included: (1) mental health, (2) physical health, and (3) social health. Several teachers focused on needs that were unrelated to content in physical education. The themes suggest that high school teachers are interested in some of the components of wellness.
Chapter Two: Aim Two

Background Information

Our grant team was provided access to the Item Bank and Testing Platform (IBTP) in order to provide a psychometric analysis of a sample of objective test items used to assess student performance in physical education, visual arts, and performing arts in K-12.

In August 2014, we received two test databases, one for physical education and the other related to various electives. The two databases contained raw pilot data from multiple tests. After searching the IBTP and following several consultations from Gary Evans at the Florida Department of Education, we determined that the first database, entitled “HtM_Elective Test Data Extract - July 2014 (1).xlsx” contained data from 11 tests. Many of the tests in this database were not available in the IBTP testing platform. However, once provided with the item identifiers, we were able to search for items from each test in the IBTP and determined that the database did not contain any data within the scope of UWF’s involvement. Specifically, the tests within this database appeared to include content related to performing work as a veterinary assistant or child educator.

The database entitled “Htm_physical Education Data Extract July 9-12 2014.xlsx” contained data for 22 physical education tests, all at the high-school level. The test names and number of participants upon which the test was piloted is listed in Table 2-1. Across tests within the database, the average number of participants upon which the test was piloted was 52.
<table>
<thead>
<tr>
<th>Test Form Name</th>
<th>Number of Participants (N)</th>
<th>Number of Test Items (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HtM_Aerobics_1_1503400_FTS14_Form#1</td>
<td>28</td>
<td>23</td>
</tr>
<tr>
<td>HtM_Basketball_1503310_FTS14_Form#1</td>
<td>54</td>
<td>25</td>
</tr>
<tr>
<td>HtM_Basketball_1503310_FTS14_Form#2</td>
<td>47</td>
<td>23</td>
</tr>
<tr>
<td>HtM_Hope Core_3026010_FTS14_Form#1</td>
<td>130</td>
<td>22</td>
</tr>
<tr>
<td>HtM_Hope Core_3026010_FTS14_Form#2</td>
<td>58</td>
<td>21</td>
</tr>
<tr>
<td>HtM_Hope Variation_1506320_FTS14_Form#1</td>
<td>69</td>
<td>37</td>
</tr>
<tr>
<td>HtM_Individual and Dual Sports 1_1502410_FTS14_Form#1</td>
<td>32</td>
<td>36</td>
</tr>
<tr>
<td>HtM_Individual and Dual Sports 1_1502410_FTS14_Form#2</td>
<td>17</td>
<td>33</td>
</tr>
<tr>
<td>HtM_Personal Fitness_1501300_FTS14_Form#1</td>
<td>116</td>
<td>38</td>
</tr>
<tr>
<td>HtM_Personal Fitness_1501300_FTS14_Form#2</td>
<td>106</td>
<td>40</td>
</tr>
<tr>
<td>HtM_Power Weight Training_1501400_FTS14_Form#1</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>HtM_Recreational Activities_1502470_FTS14_Form#2</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td>HtM_Team Sports 1_1503350_FTS14_Form#1</td>
<td>126</td>
<td>34</td>
</tr>
<tr>
<td>HtM_Team Sports 1_1503350_FTS14_Form#2</td>
<td>24</td>
<td>36</td>
</tr>
<tr>
<td>HtM_Team Sports 2_1503660_FTS14_Form#2</td>
<td>57</td>
<td>32</td>
</tr>
<tr>
<td>HtM_Volleyball_1_1505500_FTS14_Form#1</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>HtM_Weight Training 1_1501340_FTS14_Form#1</td>
<td>70</td>
<td>31</td>
</tr>
<tr>
<td>HtM_Weight Training 1_1501340_FTS14_Form#2</td>
<td>106</td>
<td>34</td>
</tr>
<tr>
<td>HtM_Weight Training 2_1501350_FTS14_Form#1</td>
<td>14</td>
<td>44</td>
</tr>
<tr>
<td>HtM_Weight Training 2_1501350_FTS14_Form#2</td>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td>HtM_Weight Training 3_1501360_FTS14_Form#1</td>
<td>21</td>
<td>32</td>
</tr>
<tr>
<td>HtM_Weight Training 3_1501360_FTS14_Form#2</td>
<td>7</td>
<td>32</td>
</tr>
</tbody>
</table>

We sought to sample the data in the database to provide an overview of the test quality. In light of this objective, we focused on tests that were broad and generalizable in content (e.g., evaluating a test related to fitness rather than one related to weightlifting). Likewise, we initially sought to sample data from various schooling levels (i.e., elementary, middle, and high school) and across the three hard-to-measure areas (i.e., physical education, visual arts, and performing arts), but were constrained by the scope of the data. We also sought to sample tests with at least 100 participants as sample size is critical to test evaluation. Thus, the sample of test forms that met our criteria used to evaluate the pilot data consisted of the following:
Analysis Results

Raw data were restructured to form three testing databases. Using answer keys from the IBTP for each of the tests, raw participant responses were transformed into zeros and ones with zeros indicating incorrect responses and ones indicating correct responses. We then examined test-level statistics including Cronbach’s alpha, means, standard deviations, and the standard error of measurement. Cronbach’s alpha is a measure of reliability that assesses the internal consistency of a test. That is, the average correlation among all of the items within a test provides an estimate of how consistently the items relate to a single construct (e.g., personal fitness). Internal consistency is impacted by how well the items capture a single construct as well as by the number of items on the test, with a tendency for longer tests to have higher alpha coefficients (Nunnally, 1967). A commonly accepted rule of thumb is that a coefficient alpha above .70 is considered good (Kline, 2000), though a much higher criterion may be recommended for high stakes testing. Because items on the tests we analyzed were dichotomously coded (correct vs. incorrect), Cronbach’s alpha and Kuder-Richardson Formula 20 (KR-20) are equivalent. Raw test means and percentage means can be interpreted as overall test difficulty assuming the questions are weighted equally.

At the item-level, we analyzed item difficulty, standard deviation, point-biserial correlation, non-distractors, and impact on the overall alpha level if the item was removed from the exam (i.e., removal of a poorly correlated item will increase the overall reliability of the test).
The item difficulty index provides an indication of whether the item is at the appropriate level or if it is too easy or too difficult. A multiple choice test with four possible responses has an optimal item difficulty \((p)\) of \(.63 - .74\) (Nunnally, 1967; Popham, 2014).

The discrimination index is an indication of whether or not the item discriminates between the lowest scorers and the highest scorers on the overall test. Minimum standards suggest that acceptable items have a discrimination index \((D)\) of at least \(.30\) (Ebel & Frisbie, 1986) when using two groups – top and bottom 50%, \(.40\) when using three groups - comprised of the top and bottom 27% (Ebel, 1954), and a point-biserial correlation coefficient criterion of at least \(.20\) for classroom exams (Wells & Wollack, 2003). As Varma (2008) recommended, we utilized the corrected correlation (adjusting the item total for the inclusion of the item being evaluated). Furthermore, negative discriminators (e.g., items for which low performers outscore high performers) should not be used. Item-level information for the three tests is reported in Appendices 2A to 2C, respectively.

Because items are coded as zeros (incorrect answers) and ones (correct answers), item means reflect the proportion of respondents who responded correctly to a given item. Point-biserial correlations reflect the relationship between each item (correct/incorrect) and overall tests scores. Non-distractors are answer choices that were not endorsed by any participants and provide diagnostic information concerning how the test can be improved, particularly to increase the test discrimination. Using participant demographic variables, race and gender, we also tested for reliable differences in overall test performance using between subjects analysis of variance and t-tests. Additionally, a subject matter expert reviewed each item to assess the quality of the wording of each item stem for such issues as bias, word length, grammar, and readability for
target grade level. The subject matter expert also reviewed the answer alternatives for similar issues, in addition to whether or not the distractors offered unintended clues to the correct answer.

**HtM_Personal Fitness_1501300_FTS14_Form#1**

The test entitled “HtM_Personal Fitness_1501300_FTS14_Form#1” contained 38 items, including two short answer questions. This test attempts to assess student knowledge of multiple strands, standards, and benchmarks documented in the Next Generation Sunshine State Standards (NGSSS). Short answer questions were subjectively graded and not reported within the dataset. Thus, our analysis for this test focused on 36 items. The average participant raw score out of 36 possible points was 14.7 (SD = 4.1). Accordingly, the average test score was 40.8% (SD = 11.4). The overall reliability of the test (alpha) was .56 (below the recommended criterion). The standard error of measurement was 7.56. Only 11 of the 38 items met the recommended threshold for point-biserial correlations.

The sample for “HtM_Personal Fitness_1501300_FTS14_Form#1” was comprised of 60 boys and 56 girls. Independent groups t-tests using participant gender as a predictor and percentage score on overall test as the outcome was not significant. That is, boys and girls performed similarly on the test. The sample contained 26 non-Hispanic Whites and 90 Hispanic/Latinos. The analysis revealed that these two groups performed similarly on the test.

**Overall Comments.** There appears to be a reasonable number of items per standard; however, both validity and reliability suffer as a result of the way in which the stems were written and the use of ineffective distractor items (see Appendix 2D). The analyses showed that
only three of the test items were within the acceptable range for optimal difficulty level (i.e., .63-.74). Using multiple methods and minimum criteria, 12 items at best meet these minimum standards (see below).

HtM_Hope Core_3026010_FTS14_Form#1

The test entitled “HtM_Hope Core_3026010_FTS14_Form#1” contained 22 items, including three short answer questions. Short answer questions were subjectively graded and not reported within the dataset. Thus, our analysis for this test focused on 19 items. The average participant raw score out of 19 possible points was 11.5 (SD = 2.2). Accordingly, the average test score was 63.9% (SD = 12.2). The overall reliability of the test (alpha) was .35, far below the recommended criterion suggesting that the test may have been assessing multiple constructs. The standard error of measurement was 9.84. Only 2 of the 22 items met the recommended threshold for point-biserial correlations meaning that test items did not accurately discriminate between top and bottom scorers.

The sample for “HtM_Hope Core_3026010_FTS14_Form#1” was comprised of 31 boys and 99 girls. Independent groups t-tests using participant gender as a predictor and percentage score on overall test as the outcome was not significant. That is, boys and girls did not perform differently from one another on the test. The sample contained 38 Whites, 73 Hispanic/Latinos, 5 Asians, 11 Pacific Islanders, 1 American Indians/Alaskan Native and 2 multi-racial students. Analysis of variance results supported that racial groups performed similarly on the test.

**Overall Comments.** Both validity and reliability suffer as a result of only including one to three items per standard (color coded on the table in Appendix 2D), using inappropriate stems,
and including often ineffective distractors (see details in table). Only one item (#14) was found to be within the acceptable range of the optimal level of difficulty. Using multiple methods and minimum criteria, only five items at best meet the minimum standards for discriminating between high and low performers.

**HtM_Team Sports 1_1503350_FTS14_Form#1**

The test entitled “HtM_Team Sports 1_1503350_FTS14_Form#1” contained 34 items, including two short answer questions. Short answer questions were subjectively graded and not reported within the dataset. Thus, our analysis for this test focused on 32 items. The average participant raw score out of 32 possible points was 16.1 (SD = 4.8). Accordingly, the average test score was 50.3% (SD = 15.0). The overall reliability of the test (alpha) was .73. The standard error of measurement was 7.65. A little more than half (19) of the 34 items met the recommended threshold for point-biserial correlations.

The sample for “HtM_Team Sports 1_1503350_FTS14_Form#1” consisted of 96 boys and 30 girls. Independent groups t-tests using participant gender as a predictor and percentage score on overall test as the outcome was not significant. Boys and girls performed similarly on the test. The sample contained 56 Whites, 37 Hispanic/Latinos, 6 Asians, 21 Pacific Islanders, 2 American Indians/Alaskan Natives, and 4 multi-racial students. Analysis of variance results indicated that race was not significantly related to test performance. However, the result approached significance ($p = .09$).

**Overall Comments.** There appears to be a reasonable number of items per standard. However, once again both validity and reliability suffer as a result of the way in which test items
were constructed. Subject matter expert comments regarding the construction of each test item is included in Appendix 2D. Only 7 of the 32 items were within an acceptable range of difficulty level. Using multiple methods and minimum criteria, 23 of the items meet the minimum standards for the ability to distinguish between top and bottom scorers.

Conclusions and Interpretation

The pilot data reviewed were limited in sample size and must be interpreted with great caution. Although there is no single gold-standard for establishing sample size, test development literature offers several “rules of thumb” and minimum thresholds. Crocker and Algina (1986), for example, stated that the sample size sought to validate the test depends upon the purpose of the test. They argued that a widely used achievement test should be validated on thousands of test examinees in a representative sample before use. In research, a commonly used standard is to seek five to ten times as many subjects as there are questions on the test (Nunnally, 1967). Based upon the low threshold of this recommendation, only one of the tests in the pilot database met the sample size minimum (i.e., HtM_Hope Core_3026010_FTS14_Form#1). The representativeness of these tests may also be called into question as it is not clear that they would generalize to the larger population of Florida students within the same grade level. As one example, no African Americans were present in the sampled data. This is particularly concerning considering the ethical and legal issues surrounding test development (e.g., see Ployhart & Holtz, 2008).

Nonetheless, we discuss and interpret the results of our analyses of the pilot data we sampled in regard to accepted standards below. A value of .70 or higher is often cited as an acceptable minimum for KR-20 and Cronbach’s Alpha in research (Crocker & Algina, 1986;
Kline, 2000) and has also been cited as a standard for classroom tests (Wells & Wollack, 2003). However, standards for overall test reliability vary by test purpose. Wells and Wollack (2003) stated that while .70 is acceptable for classroom tests, a reliability between .80 and .85 is acceptable for standardized low stakes tests; a high-stakes standardized test should have a reliability of .90 or higher (Wells & Wollack, 2003). Thus, while HtM_Team Sports 1_1503350_FTS14_Form#1 had a reliability above .70, it is not adequately internally consistent for standardized testing. A review of the last column of the appendices within this report will indicate the low internal consistencies cannot be attributed to any particular item. As discussed above, many of the items on the tests failed to correlate with overall tests score, not meeting the overall threshold for point-biserial correlations of .25 or higher (Varma, 2008) or even the more liberal criterion of .20 or higher used in the current study (Wells & Wollack, 2003). Indeed, some of the items in the appendices exhibited negative point-biserial correlations indicating that getting the item correct was associated with a lower test score. In such cases, items should be carefully examined. For example, they may have more than one plausible correct answer, elicit multiple interpretations, or be miscoded as correct.

The mean percentage scores for each of the tests were low. Because means and standard deviations are impacted by the scale on which they are measured and the purpose of the test, there is no literature-based benchmark to use as a point of comparison. If percentage scores are interpreted as course grades, across tests, the average participant failed the test (mean < 64%). Tests tend to be more reliable when the difficulty value of items is mid-range (.5 and above; Office of Measurement and Evaluation of Teaching, 2015). However, as reported in the appendices, many items failed to meet this threshold. Students were largely unfamiliar with test content.
In light of the low sample sizes, limited demographic composition of the sample, and the performance of the tests in relation to test development standards, we recommend that the testing database be carefully evaluated and re-piloted before it can be used to make educational decisions. In addition to the analyses just described, subject matter experts provided a qualitative assessment of the test items. Subject matter experts’ independent qualitative analyses in large part, corroborated the psychometric analyses. That is, many of the items were found to include irrelevant information, provided unnecessary or unintentional clues to the correct answer, grammatical errors, or were judged to be potentially confusing to the intended test taker. Though we only analyzed a sample of three tests from the test bank, none of these tests should be considered as possessing the psychometric properties necessary to make informed decisions regarding student growth in the discipline of physical education. Please refer back to Volume I for recommendations in conjunction with the larger proposed model.
Chapter Three: Aim Three – Input from FL Stakeholders

The third aim of this project focused on identifying performance measures that may be used to assess music education, visual arts education, and physical education. Toward this end, we reached out to district-level subject matter experts within the state of Florida to identify existing measures and searched both the academic and practical literatures. This chapter focuses on our inquiry with FL district-level subject matter experts. The next chapter covers the literature review.

Method

Design

Researchers gain in-depth information from focus groups by enlisting experts in a given field to answer structured, open-ended questions to facilitate organic discussion. Online focus groups have been more recently utilized due to their low cost, potential to reach a geographically diverse audience, and ease of access for busy professionals (Edmunds, 1999). Thus, we created two independent, qualitative discussion board using BlogNog, a secure, asynchronous online discussion board. Computer-mediated focus groups, as compared to face-to-face, generate less communication between participants, but produce more novel ideas and answers to structured questions (Gallupe, Bastianutti, & Cooper, 1991; Reid & Reid, 2005). Online communication also enables anonymous responses and thus promotes honest responding while lessening the risk of adverse consequences for sharing these opinions. Another benefit, encouraged by the question instructions, is the ability to attach lesson plans or supporting documentation. For the purposes of collecting responses to a sensitive topic, extensive participant interaction was not
critical to the desired outcome. We aimed to gather in-depth responses to the structured questions.

This portion of the project was approved by the Institutional Review Board at the University of West Florida (IRB 2015-129). See Appendix 3A for the approval. Participation was voluntary. The information linking individual responses to participant names was kept confidential.

Participants and Recruitment

Our subject matter experts used purposeful, stratified sampling to identify curriculum administrators and current teachers in their respective subject areas across the state of Florida. This approach is similar to previous recruitment strategies for online focus groups by taking advantage of easily definable population groups, and follows recommendations to increase response rates by using existing personalized contacts (Cook, Heath, & Thompson, 2000; Witmer, Colman, & Katzman, 1999). In music education, three administrators and eight teachers were identified. In visual arts education, six administrators and five art teachers were identified. In physical education, six administrators and six teachers were identified. These potential participants were then contacted by the subject matter expert and agreed to participate. The initial email invitation (Appendix 3B) was sent on March 10, 2015 with reminder emails (Appendix 3C) sent on a weekly basis until the discussion board closed on April 10, 2015.

We were able to achieve a stratified sample that was representative of teachers of the disciplines and contained both educators and administrators. In prior research, focus groups have remained active for up to two months, and response rates are still traditionally low, ranging
from five percent to 20.7% (Williams, 2003; Witmer et al., 1999). Comparatively, our response rates were very good. A total of 11 teachers participated in the online focus group, resulting in a response rate of 58%. Four administrators participated, resulting in a response rate of 27%.

The teacher subsample is representative of each of the three hard-to-measure content areas: music (45%, \( n = 5 \)), visual art (27%, \( n = 3 \)), and physical education (27%, \( n = 3 \)). The participants averaged nine years in their current position (SD = 7.95) and 15 years overall as a classroom teacher (SD = 7.64). Thirty-six percent of participants were men (\( n = 4 \)), and 64% were women (\( n = 7 \)). Participants taught various grade levels: K-5 (27%, \( n = 3 \)), K-8 (18%, \( n = 2 \)), 6-8 (9%, \( n = 1 \)), 6-12 (18%, \( n = 2 \)), and 9-12 (27%, \( n = 3 \)). A variety of counties/schools were also represented, such as Escambia (64%, \( n = 7 \)), Florida Virtual School (18%, \( n = 2 \)), Orange (9%, \( n = 1 \)), and Florida School for the Deaf and Blind (9%, \( n = 1 \)). Of the four administrator respondents, two participants failed to provide complete information to be retained in the sample. To protect participant anonymity, limited participant demographics are reported. Both administrators worked in their position for eight months; one taught grades K to 8 and one taught grades K to 12.

Procedure

One BlogNog discussion board was created for participation by supervisors, and a separate BlogNog discussion board was created for teachers at various grade levels in music, art, and physical education. Both discussion boards included four areas: Informed Consent and Demographics, Music Education, Art Education, and Physical Education. The Informed Consent and Demographic questions can be found in Appendix 3D. The important skills for each content area from Aim 1 were included to gather further information on how to assess these specific
skills (Appendix 3E). Then, seven identical questions were developed for each discussion board (Appendix 3F). An example question includes, “Based on your experience and/or expertise, how do you recommend that teachers assess performance in the skills listed above?”

Participants were sent an email with the unique link, and were informed that participation would take around 20 minutes. They were asked to monitor the discussion board regularly. The purpose of these instructions was to promote a discussion similar to a traditional focus group format. BlogNog allows anonymous responding by assigning participants a unique link and number. Participants were identifiable by their number, which accompanied their discussion board responses. Anonymous responding during web-based discussion boards allows researchers greater control of protecting identity information and ensuring responses are kept confidential over face-to-face focus groups (Stewart & Williams, 2005). Additionally, other participants could not see responses to the Informed Consent and Demographics, but were able to openly view other responses on the remaining discussion boards. A co-PI and graduate research assistants frequently monitored the responses and asked follow-up questions when necessary.

Results

Overall, most district-level administrators and in-service teachers agreed on two main points: growth and proficiency should factor into a multiple assessment system, and at least 50% of the evaluation should be based on performance. Knowledge was either equally or slightly less weighted, while appreciation was not generally considered to be a factor that is difficult to measure. Also, some suggested that more emphasis should be placed on growth in elementary grades, because proficiency is more related to the skill development in upper grades.
Music Education

Due the variability in performance mediums (e.g., choral versus instrument), flexibility was an area of focus in this content area. The current practices described are inherently flexible because selection is mostly left to the teacher, and participants listed rubrics, self-assessment, and peer assessment as common tools. One participant mentioned the use of SmartMusic software and seemed to have a positive response to its usefulness, but another noted its failure to assess tone quality. A “pass-off” system, in which a student must master one skill before moving on to the next, was also described. From those who were involved in the formal Music Performance Assessments, appreciation for feedback from a panel of experts was expressed.

Regarding the more subjective skills such as risk-taking and respect for the creative process, one participant confessed a lack of understanding for how to rate achievement in these areas. This concern was also reflected in other responses. Such feedback is a valuable insight to the need for professional development with future implementation of assessment procedures.

Visual Arts Education

Participants seemed displeased with the inconsistency of assessment methods currently being used, and suggested using approaches that could incorporate performance and cognitive (e.g., multiple-choice) content for end-of-course exams. Even one educator who was opposed to art assessment for elementary grades was open to the use of the Florida Art Education Association (FAEA) rubrics as well as the collection of performance artifacts (e.g., pieces of artwork; see Appendix 3G). The FAEA rubrics were also encouraged as a method for portfolio
evaluation by a district administrator. Digital storage of artwork (via images) was strongly preferred over hard copy, given the typical number of students in art classes.

Physical Education

District-mandated wellness plans were well liked by the participants who discussed them. Students apply principles (MAD, FITT, and SPORT, all related to effective training) to develop a personal fitness program. They use wearable technology (“movbands”) to track and record their progress, which is recommended over paper-and-pencil tracking because of class size and remote accessibility (e.g., the students can practice at home, which would address one of Florida’s outside-of-school activity benchmarks).

Portfolios, rubrics, and checklists were also named as teacher-selected assessments that could generate performance data. However, one participant indicated the relationship between integrity and data collection as being susceptible to external pressures when teachers are tasked with inputting their own student scores. Finally, it should be noted that special consideration for students with disabilities was encouraged for this particular content area.
Chapter Four: Aim Three – Literature-Based Inquiry

This chapter covers the literature-based inquiry portion of Aim Three, which sought to identify performance measures. The purpose of this chapter is to serve as a reference; it is not intended to be read in a single sitting. As we evaluated potential measures, we considered both advantages and disadvantages. Among the advantages and disadvantages considered were reliability, validity, practicality/logistics, cost, precedence and anticipated stakeholder acceptance. Based on our review of the literature, we categorized performance measurements into the following groups: (1) Recorded performance assessments, (2) Student logs, (3) Sketchbooks, (4) Fitness plans, (5) Self-assessment/reflection, (6) Fitness tests, (7) Checklists, and (8) Rubrics.

(1) Recorded Performance Assessments

Music Education

A general approach in hard-to-measure content areas is to require a sample recording generated independently by the student, such as an audio file of a musical piece, a digital image of a sculpture, or a video of a dance performance (Goolsby, 1995). This is submitted much like a paper assignment would be and is graded by the teacher. With this method, long-term access should be considered (e.g., devices that play CDs are much less common than they used to be; Cennamo, Ross, & Ertmer, 2014). These recordings take different forms across subject areas, but the sound, image, and motion needed for music, art, and physical education may also be integrated in a single product (e.g., captured by camera, presented in Microsoft PowerPoint). The main factor to consider in choosing software is whether it should include assessment and interactive capabilities on top of being a recording device. Software used to
record a music performance is known in the field as a “digital audio workstation” (DAW) and has a wide range of price and quality. Education-focused packages can include the evaluation component, as well. Criswell (2015) cited high school music teacher Richard McCready as making the comment that in the classroom, the most important feature is that the interface is not intimidating to students and that it can be used effectively for teaching.

**Visual Arts Education**

In visual arts, electronic images and hard copies of the same artwork have been shown as having intra-rater reliability (Dorn & Sabol, 2006). Images, as opposed to actual pieces, also support portfolio assessment since advantages to this method of storage include accessibility by others and portability (Siegle, 2002; Sunstein & Lovell, 2000 as cited by Dorn & Sabol, 2006). The impact of technology in this field is so widespread that even professional galleries utilize digital images (Poole & Ho, 2011). Dillon, Nalder, Brown, and Smith (2004) discussed how electronic formats can “improve the scope, depth, relevance, and frequency of feedback in arts assessment” in addition to increasing rigor and accountability (p.37).

**Physical Education**

The applications of technology in physical education are less intuitive, but important to consider nonetheless. In fact, the NGSSS standards for physical education include a benchmark for the use of technology in fitness, which specifically mentions Wii Fit and Dance Dance Revolution software. Tannehill, van der Mars, and MacPhail (2015, p.66) discussed schools that have utilized these video games along with online communication to monitor a student’s fitness levels electronically. Other studies have investigated pedometers and heart rate monitors, finding that these tools were a valid indicator of performance levels in students (Patridge, King,
Similar to music assessment software, these games can provide informal assessment in addition to fitness records.

**Music Education: SmartMusic®.** SmartMusic® is an interactive tool to facilitate performance assessment between teachers and their classes (MakeMusic Inc., 2015). Teachers can select a task from the program’s library or create their own. Once an assignment is created, it can be sent out to the class or specific students to use at school or at home. The software displays the student’s performance in the form of written music, and the correct notation appears in a different color (see Figure 4-1). A percentage of correct notes is generated at the end of the piece. In an experimental test of this assessment, Long (2011) noted the dichotomy of grades assigned (e.g., mostly A’s and F’s). In this way, it cannot replace a human evaluator who is able to take contextual factors into consideration and serves as a primarily pass/fail assessment. However, students can use multiple attempts to rehearse and select their best performance to send to their teachers. If teachers want to consider other factors besides correct notes, they can listen to the recording and change the grade in the software’s gradebook manually, leaving personal feedback for the student. A particularly useful feature of this software is that state standards can be attached to a task, allowing for easy generation of a standards-based portfolio artifact.

Costs associated with this program are a teacher subscription, student subscriptions, and equipment such as an iPad for using the program and a microphone for recording work. It costs $140/year for educators, $44/year per student, or $44/year for a practice-only version that both teachers and students can use to rehearse but not to record. Though slightly cheaper software packages for recording music exist (e.g., Finale), SmartMusic® was designed specifically for
young people and classroom needs. Other packages with this design are more expensive, e.g., Alfred’s Essentials of Music Theory, which costs $60 per student. Comparatively, it is cost effective for this type of DAW and would benefit large, performance-focused classes, but could still be too expensive to require statewide.

Figure 4-1: Screenshot from SmartMusic® Recording and Assessment Tool (MakeMusic Inc., 2015). This program records student performance and provides a ratio score of correct notes. With this tool state standards can be attached to a task, allowing for easy generation of a standards-based portfolio artifact. Copyright © 1993-2012 MakeMusic, Inc.

Visual Art Education 2: Digication. At $34.95/year for teachers and $20/year for students, Digication is a slightly less expensive electronic recording package for visual arts. This web-based software allows students to create multimedia e-portfolios using a variety of templates. It mimics a personal website, which can be constructed at intervals and appears easy to navigate (see Figure 4-2). Similar to SmartMusic®, the teacher can communicate with students and assess their work through the program. Digication is currently used in many
colleges and public schools for art assessment. Because of its audio and video capabilities, it could be considered for use in all three content areas.

Figure 4-2: Screenshot from Digication e-Portfolio construction (Digication Inc., 2015). This web-based software allows students to create multimedia e-portfolios using a variety of templates, similar to a personal website. Copyright © 2015 Digication, Inc.

Physical Education: Wii Fit™ and Wearable Fitness Technologies. Makers of video game consoles have targeted physical education programs by offering all-inclusive packages designed to teach and measure activity. Depending on the brand and console, the system may use a handheld remote, weight-sensitive mat, or scanning “eyes” to assess the body position and physiological status of the user. For example, the Wii Fit™ for Kids equipment can be purchased based on class size (e.g. $995 for 25 students + teacher) and includes balance boards that measure weight and register pressure. The Alabama Department of Education currently uses Wii Fit™ in 70 schools three times a week, and has plans to continue expanding the funding program (Wee Can Fight Obesity) until all Alabama schools are participating (PowerPR, 2011). Figure 4-3 shows a screenshot from Wii Fit™.
As a final note, there is an increasingly large variety of “wearable technologies” that record physical activity and include data tracking software. For example, the FitBit® bracelet (approx. $129 per person) is worn on the wrist and can be connected to a computer to upload and store physiological data (see Figure 4-4). In Florida, focus group participants indicated the use of a similar product with a smaller price tag ($30/per student) that is used in some districts to track activity both in and outside of school (“Movband®,” see Figure 4-5). Empirical research is needed to ascertain the accuracy of these products, since consumers offer mixed reviews.

Figure 4-2: Screenshot from Wii Fit™. Similar to a video game console, this software can be used in classrooms (e.g., Alabama Department of Education has funded it in 70 schools) with accessories such as weight-sensitive balance boards. Copyright © 2008-2011 Nintendo.
**Figure 4-4:** Screenshot from FitBit® Data Tracking Website (FitBit Inc., 2015). By connecting an enabled wristband to a computer, this software tracks and stores personal fitness data.

Copyright © 2015 FitBit Inc.
Figure 4-5: Movband® Fitness Tracker (Movable, 2015). Used in some Florida districts, this wristband is worn to track physical activity inside and outside of school. Copyright © 2015 Movable LLC.

(2) Student Logs

Literature is mixed on operationally defining a student “learning log” (Northern Illinois University, N.D.). For our purposes, student learning logs are journal entries recording the learning process, questions asked while learning, and/or time spent practicing activities (Equipped for the Future, 2004; West Virginia Department of Education, n.d.). Learning logs create an ongoing dialog between student and teacher, which is used for evaluation and also allows for the teacher to provide feedback to the student as needed (Sulzberger, 2014).
Student logs can be beneficial in a variety of subject areas including the arts and physical education (e.g., Chen, Mason, Hypnar, Zalmount, & Hammond-Benett, 2014; Dolmans, Schmidt, Beek, Beintema, & Grever, 1999; Pembrok & Fredrickson, 2001; Tornoe, 2007). Student logs are a form of reflective learning and provide a greater sense of independence as well as a consciousness to the learning process itself (Stephens & Winterbottom, 2010). Sulzberger (2012) also suggested that student logs are effective in providing feedback opportunities; the formative feedback provided in a timely manner can make a positive difference in the student learning outcomes. This may suggest that waiting too long to give feedback to students on performance may not be as beneficial to learning outcomes. The use of student logs in education allows the student to engage in a form of self-directed study, which has been shown to be desirable to students (Rhoads & DeHaan, 2013). In support of this notion, there is some evidence to suggest that the content of learning logs changes over time, becoming increasingly complex (e.g., Klemp, 2010). Rhoads and DeHaan (2013) suggested that teacher interactions play a large role in student success in self-directed study. Thus, feedback in student logs can be crucial to the completion of learning outcomes.

Music and Visual Art Education

Oreck, Owen, and Baum (2003) stressed the importance of authenticity when considering the validity of art and music assessment. In their words, “if the artistic experience is not authentic, then the students' responses are unlikely to be artistic” (p.67). They suggested this is achieved through engaged participation over multiple sessions, which was also identified as important by Florida educators. For engagement to be recorded and included as a portfolio artifact, logs of rehearsal time and class participation would be needed. While they are not often
considered as a stand-alone method because they cannot offer achievement data, student logs can function as a component in comprehensive assessment by providing documentation and material for reflection.

Darling-Hammond (1994) identified the need for transparency in evaluation to unite students, teacher, and administrator goals. Additionally, she recommended that both process and product should be included in a comprehensive assessment. While many sources concur with these ideals, few offer solutions in the form of concrete examples (Bensur, 2002; Fisher, 2008; Russell & Austin, 2010; Scott, 2012). Time-based logs are one way to provide clear and sequential evidence of participation, which necessarily precede an artistic creation. They can also support reflective assignments by allowing the student to analyze the relationship between process and product. For example, a student who has logged rehearsal time on a certain technique can later write a reflection on whether this time was adequate for the intended product.

Physical Education

Student logs are commonly used as records of student engagement in physical activity outside of the school day in physical education programs (Chen et al., 2014; St. Ours & Scrabis-Fletcher, 2013). In this way, they are considered a form of “active homework” and typically require a guardian’s signature (St. Ours & Scrabis-Fletcher, 2013, p.23). To address performance within the school day, these records can be integrated with others in order to accomplish multiple purposes.

Paired with reflective questions, Gibbons and Kankkonen (2011) suggested physical activity logs as evidence of self-assessment that can be used during class time. They can be used
as a means to record the amount of time and time of activity completed. Then, the students can evaluate themselves using prompts that generate evaluative responses. The rationale is that students with access to personal data can draw conclusions about their progress to develop “internal feedback” (Gibbons & Kankkonen, 2011, p.8). An example of this pairing is an end-of-week check-in torecord and analyze the effects of a particular activity. Elliott (2004) also suggested the use of logged activity to inform a personal fitness plan, which coincides with benchmarks selected for assessment in the current study.

Logs can also help reconcile some of the limitations to performance assessment in physical education. For instance, tracking personal activity is shown to be more motivating for students who work from a task-oriented perspective as opposed to those who respond to competitive conditions, according to goal-orientation theory (Weirsma & Sherman, 2008). Including activity logs would allow these types of students to show participation without the potentially stressful effect of comparison to others. Conversely, students who are working on specific skill development in competitive sports can use the data differently (Weirsma & Sherman, 2008). This flexibility presents challenges in assessing the reliability and validity of activity logs, but including such an artifact has the benefit of authentic, individual representation.

(3) Sketchbooks

A sketchbook (also called idea journal, visual journal, or visual thinking journal) is a notebook in which students can combine written and visual imagery to promote reflection (Deaver & McAuliffe, 2009). Anderson and Milbrandt (2005) described a student’s sketchbook as a place to consider topics and apply critical inquiry, though they can be used for different purposes such as focusing on one subject or recording impressions from a range of topics.
(Christensen, 2010). Depending on the objective, contents may include rough drafts, media clippings, textile samples, and other signs of artistic process. Visual artifact journals have also been used to promote student reflection in which the student included a personal artifact (e.g., ticket stub, photograph, plants, labels, jewelry) and wrote a descriptive, critical, or reflective response (Sanders-Bustle, 2008).

Parallel with the recommendation of process portfolios, sketchbooks may be used to document and assess a student’s progress and mastery. The sketchbook is submitted periodically and evaluated with a rubric (Morris School District, n.d.). Clear standards, support, and criteria should be in place to facilitate grading and encourage technical and intellectual development (Froslev, 1994). In light of this, one drawback of submitting sketchbooks for review is the limitation of student expression due to lack of privacy. However, they have the benefit of being an engaging medium to promote and track student progress and are applicable to diverse students (Froslev, 1994).

(4) Fitness Plans

Schools are tasked with promoting both physical fitness and health initiatives. To this end, students in health and physical education classes are often required to develop fitness plans as a part of course requirements. A fitness plan can provide a means by which one can achieve a specific fitness goal (Ayers & Sariscany, 2010). Regardless of the goal, conditions exist for developing suitable plans. For example, seven primary elements have been identified as important for inclusion in fitness plans: cardiovascular and aerobic conditioning, strength training and muscular development, stretching, core stability, nutrition and supplementation, and
mental rest and relaxation (Ayers & Sariscany, 2010). Additionally, plans should be created with the needs of the individual in mind so specific goals can be met (Su, Chiang, & Chih, 2014).

Generic fitness plans are readily available online; however, these plans may not take into account these seven primary elements or an individual’s specific needs, goals, strengths, and limitations. Some websites allow users to develop a more personalized fitness plan (see Figure 4-6). While plans can vary from general to specific for an individual, these plans often vary considerably in both detail and price (Workouts for You, 2015). Nonetheless, it should be noted that customized plans are more likely to be effective and enjoyable since the individual is personally involved in its development (Wiersma & Sherman, 2008).

Mercier and Silverman (2014) suggested that fitness activity should be a critical component of fitness instruction and that students should use self-assessment to show improvement. In alignment with this, fitness plans in physical education offer a formal type of self-assessment while addressing each of the three areas of the model (i.e., performance, cognitive, and affective). Specifically, they demonstrate the ability of a person to self-assess students’ own needs and preferences and then create a plan that fits into their current lifestyle and ability. The fitness plans can be submitted as course artifacts that can be evaluated for effectiveness, objectives covered, and goals met (Welk, 2008).

Indeed, a student-centered fitness plan provides an authentic assessment that allows for students to engage in meaningful activities that can enhance both knowledge and participation (Kearny, 2013). Grading in physical education should not be based on fitness level, but instead on the ability of students to assess their own fitness level and create goals that will help guide them to healthier lifestyles (Feith, 2014). When taking this into consideration, it becomes a
strong argument for inclusion into a portfolio. Students can complete a pre-test before they create a personalized fitness plan and then follow up with a post-test once they have met their goals, providing relevant information for inclusion in a portfolio (Feith, 2014).

Student-created fitness plans can provide physical education teachers a way to assess several different areas and benchmarks. Based on teacher-created lessons that address pertinent health, skill, and physical activity components, students would be able to develop fitness plans and then self-assess via their fitness plans. Furthermore, the creation of a fitness plan gives the instructor a way to assess the cognitive ability of the student in relation to the fitness lessons. Additionally, the pre-test allows students to self-assess their current fitness levels. Finally, student participation in the fitness plan allows the instructor to assess student performance. Thus, each component of the model (cognitive, performance, and affect) can be assessed. Consequently, it is suggested that rubrics address goal setting and planning (cognitive), implementation (performance), and evaluation (affective) be used. It should be noted that reliable (above 70% agreement) rubrics for assessing student fitness plans exist that allow multiple raters to rate the fitness plans (Day, 2000; Urban & Archibald, 2013).
Figure 4-6: Sample Weekly Fitness Plan and Daily Exercise Log. A student can create a goal-oriented plan for activities and combine it with record-keeping to assess adherence to the plan (University of Missouri, 2015).

(5) Self-Assessment/Reflection

The practice of self-assessment is a useful life skill for students (Hanrahan & Isaacs, 2001). Student self-assessment can be used to encourage reflection on both the final product and the process of creating (Gitomer, Grosh, & Price, 1992). At the core of both the teacher/student portfolio interaction and student self-assessment is questioning. Questions that encourage self-assessment and self-reflection tap into the responding aspect of the arts, which is linked with the affective component of learning, and provide insight to student learning (Gitomer, Grosh, & Price, 1992; Winick, Avallone, & Crovo, 2008). Self-assessment is widely used in the classroom and shows merit in increasing student achievement and improved behavior (Ross, 2006). Thus,
when we talk about affective assessment, this may refer to teacher-ratings of student attitudes or self-assessment of how personal attitudes, values, or responses have changed with regard to the particular learning outcome.

**Reliability and Validity of Self-Assessment**

The internal consistency reliability of self-assessment is high, and ratings appear stable in different subject areas, but less stable over time (Blatchford, 1997; Ross, Rolheiser, & Hogaboam-Gray, 2002; Sung, Change, Chiou, & Hou, 2005). The practical implication is that self-assessment is a reliable measure, but young children may be less effective at producing accurate self-assessment ratings consistent across time and content area.

Validity evidence has mixed results. A meta-analysis showed an acceptable correlation between teacher and student assessment scores (Ross, 1998). Differences in ratings between student and teacher raters may be due to over-inflation by the student, especially in younger children, or student lack of knowledge regarding the assessment domain (Butler, 1990; Ross, 1998). Teacher led self-assessment training for students increases the validity and reliability of self-assessments (Hanrahan & Isaacs, 2001; McDonald & Boud, 2003).

Researchers have discussed two possible ways to standardize the measurement of affective learning through self-assessment. Gibb’s (1988) reflective cycle is a popular model for guiding student reflection. This approach provides the student with guidance for which information to incorporate into future attempts at meeting the learning objective. The author recommends six steps, which can be applied to any learning experience (see Table 4-1).
Table 4-1. Gibb’s Reflective Cycle (1988)

<table>
<thead>
<tr>
<th>Step</th>
<th>Domain</th>
<th>Reflective Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Description</td>
<td>What happened?</td>
</tr>
<tr>
<td>2</td>
<td>Feelings</td>
<td>What did you think and feel about it?</td>
</tr>
<tr>
<td>3</td>
<td>Evaluation</td>
<td>What were the positives and negatives?</td>
</tr>
<tr>
<td>4</td>
<td>Analysis</td>
<td>What sense can you make of it?</td>
</tr>
<tr>
<td>5</td>
<td>Conclusion</td>
<td>What else could you have done?</td>
</tr>
<tr>
<td>6</td>
<td>Action Plan</td>
<td>What will you do next time?</td>
</tr>
</tbody>
</table>

Another approach is Johns Model for Structured Reflection (Johns, 1995), which may be used in a journal in any hard-to-measure area to carry out affective self-assessments. This model guides students to look inward (e.g., focus, write down thoughts and emotions) and outward (e.g. What was I attempting to achieve? What factors influenced me? Does the situation connect with previous experiences?). These models are highlighted due to their practical use for devising a standardized assessment for affective reflections across the state. Next, we will continue to discuss the use, tools, and examples of affective assessments in music, visual arts, and physical education via teacher- and self-assessments.

Music Education. Parallel to visual arts education, the NCCAS standards recommend assessing performing arts students’ ability to create, perform, and respond. To do so accurately, examples from their collaborative Model Cornerstone Assessments (see example in Figure 4-7) include self-assessment and reflection components. These are used in reference to both performance and appreciation of musical works.
Check any categories that make a connection for you with music.

<table>
<thead>
<tr>
<th>Specific Interest</th>
<th>Experience</th>
<th>Purpose</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music style</td>
<td>Heard it before</td>
<td>Ceremony</td>
<td>Cultural</td>
</tr>
<tr>
<td>When written</td>
<td>Performed it before</td>
<td>Recreation</td>
<td>Historical</td>
</tr>
<tr>
<td>Elements used</td>
<td></td>
<td>Commercial</td>
<td>Personal</td>
</tr>
<tr>
<td>Emotional</td>
<td></td>
<td>Expression</td>
<td>Social</td>
</tr>
</tbody>
</table>

Explain why you selected this connection(s) to Simple Gifts:

_______________________________________________________________

Figure 4-7: Music affective self-assessment. This example was provided by the NAfME recommended Music Model Cornerstone Assessment for the Artistic Process: Responding, 5th Grade General Music, Question 8, p. 7 (2015). This example prompts students to self-assess their attitudes toward a song.

Mills (2009) wrote about both learning portfolios (progress) and best-works portfolios (proficiency) and made recommendations on incorporating student self-assessment in learning portfolios. An adapted portfolio rubric is presented in which students work on a “musical goals” worksheet and reflect on the process of creating musical goals (see Appendix 4A). A rubric such as this may be used to evaluate the portfolio as a whole based on a specific sequence and product. It can be paired with a weekly class schedule that ensures timely completion of artifacts (e.g., Week 1 = Reflection 1). Prompts for reflective writing include “Write a short statement reflecting your experience writing musical goals. What was easy, difficult, surprising, frustrating, energizing and/or interesting about this experience?” (Mills, 2009, p. 35).
Recommended by the national standards and utilized in classrooms, affective learning is an important component of learning in performing arts education, which can be assessed via the use of self- and teacher-assessment. Due to the nature of performing arts, videos are a useful tool for teachers to promote review and reflection of student performances (Information Resources Management Association, 2012). A drawback of recording students’ performances may be the acquisition of technology and time it takes to review the recordings, so other options are available. For instance, journal entries can serve as an artifact of a student’s affective self-assessment. By examining the quality of the works they produce, reflection may be directed towards goals such as appreciating strengths and strategizing to improve weaknesses.

Individual differences will factor into the engagement and assessment of journal entries (Social Policy and Social Work, n.d.). Some caveats are that students may be inclined to write minimally, write what the teacher wants, or develop antagonism when they do not understand the learning goal (Finlay, 2008). The addition of structured questions to help students engage in self-assessment is useful to guide affective response efforts (Social Policy and Social Work, n.d.). Finlay (2008) recommended and expounded upon the following guiding principles to promote effective reflection practices in students:

- **Present reflective practice(s) with care:** Motivate students, provide students with different models of reflection, use formative feedback instead of assessment in early stages of reflection.

- **Provide adequate support, time, resources, opportunities and methods for reflection:** Students should practice self-reflection in private and in dialogical team contexts, as well
as reflecting on different mediums (e.g., case studies, their own work, performances, critical incidents).

- **Develop skills of critical analysis**: Introduce different ways of critical thinking and questioning (e.g., ethnomethodological, hermeneutic phenomenological, and discursive analysis).

- **Take proper account of the context of reflection**: Student reflection occurs in the context of the institution, political, and cultural climate and encouraging questions and critical thinking about these influences will promote more in-depth reflections.

**Visual Arts Education.** The National Assessment of Educational Progress (NAEP) created the Arts Education Assessment Framework as a guideline for assessment in various artistic disciplines (Winick, Avallone, & Crovo, 2008). In this framework, evaluating the affective component is suggested in accordance with the national standards, which included responding to visual art (NCCAS, 2013). Visual art is similar to music in that self-assessment is also supported by affective tools such as reflective journaling with prompts.

Examples of prompts include asking students what they have learned, what they would change about a specific classroom activity, and what they found most/least interesting about it (Education Services Australia, n.d.). In the International Baccalaureate (IB) Art/Design program, self-evaluations using this approach were conducted four times a year using a written form. The form included similar questions about the interest and inquired about the use of design principles for expression (Tomhave, 1999). Responses such as these can be used with portfolios as both an artifact and a general self-assessment using data collected by the portfolio itself (see Figure 4-9).
Instructions: Looking back through your portfolio of artwork take this time to reflect on your development as an artist. Take all of your works of art into consideration as you respond to the questions below. Pay close attention to your sketchbook assignments as well. Your opinions perceptions and thoughts are important to your development as an artist. Use complete sentences and artistic terminology in your responses.

1. Which characteristic of style can you identify as uniquely your own? (color choices, use of line, brush control, construction skills, etc.) Explain.
2. Which piece of work is your favorite? Explain why.
3a. Select a piece that you are displeased with. Why are you displeased with the piece?
3b. What did you learn from this piece of artwork?
4. Select a piece of art work. How would you alter or change this piece if you were to do it again? Explain.
5. What growth do you recognize in your artistic skills since beginning this semester?”

Figure 4-9: Reflective prompt. This example was used as part of a final exam in a high school arts classroom from Gitomer, Grosh, and Price (1992).

Rubrics are also used in conjunction with prompts in affective domains. One of the structured questions provided in the NAEP (2008) framework discusses features of a self-portrait that express personality, and the grouping of portraits according to mood/feeling is later evaluated using a rubric (see Figure 4-10). Similarly, students in the IB program judged themselves on a holistic rubric with a 1-5 scale on several dimensions, which they also had to justify in writing (Tomhave, 1999). Teachers completed the same form and responded to the students’ comments. These examples provide evidence of the widespread use and utility of affective assessment via self and teacher ratings in visual arts education.
<table>
<thead>
<tr>
<th>Brief Description of the Assessment/Key Traits</th>
<th>No evidence</th>
<th>Limited evidence</th>
<th>Sufficient evidence</th>
<th>Strong evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students collaboratively examine and respond to a diverse selection of self portraits (Responding)</td>
<td>Student did not participate in the examination and discussion of self-portraits.</td>
<td>Student examined the artwork but did not respond to any questions or comments made about the self-portrait.</td>
<td>Student examined the artwork and made limited contributions to the discussion.</td>
<td>Student examined and responded to a diverse selection of self-portraits, making contributions in both areas.</td>
</tr>
</tbody>
</table>

Figure 4-10: Visual art self-portrait presentation rubric. This example was provided by the National Coalition for Core Arts Standards Visual Arts Model Cornerstone Assessment: 2nd grade, p.16 (2013). This task is completed by teachers using a 4-point scale from “no evidence” to “strong evidence.”

Physical Education. National and state standards feature several skills that align with self-assessment and reflective tasks for physical education (National Association for Sport and Physical Education, 2011). Kaardal (2001) suggested applying self-evaluation specifically in the area of fitness since it naturally requires individualization. Objective grading can be established ahead of time by giving students the option of comparing their scores to age-referenced norms or personally-selected targets.

Self-assessment and reflection methods are versatile in physical education and can be adapted from a range of the other types discussed. Checklists of their own behavior (e.g., proper warm-up and cool-down), fitness/nutrition journals that reflect on habits, and performance rubrics with recorded video or activity data are recommended by multiple sources (Corbin & Lindsey, 2002; Kaardal, 2001; Thomas, Lee, & Thomas, 2008). While teacher- and peer-
observations provide the opportunity for students to receive feedback on their progress, written self-assessments allow them to voice their reflections and examine performance independently, which is a lifetime fitness skill (Corbin & Lindsey, 2002). These tools are also useful for identifying activity preferences and effects, which coincide with current benchmarks.

(6) Fitness Tests

Youth fitness tests have been used as an assessment tool in physical education to help students understand their fitness level for many years (Keating & Silverman, 2009; Liu, 2008). Two of the major tests, the President’s Challenge and Fitnessgram, were recently combined and retained the Fitnessgram name. This commonly used test focuses on the health-related components of fitness (cardiovascular endurance, muscular strength, muscular endurance, flexibility, and body composition), which are also ingrained in current national and state standards. For reference, measures of these components are presented in Table 4-2.

There are serious concerns with using the tests as a stand-alone measure for physical education. The content of these national fitness tests are only a portion of the material covered in physical education classes, so they are not able to address a school year’s worth of learning outcomes (Keating, & Silverman, 2004; Mercier & Doolittle, 2013). Beyond issues of content validity, they have been subject to scrutiny for issues of accuracy and impact. Scholars have questioned the generalizability of a “snapshot of a child’s performance on a single day, decontextualized form the circumstances of the student’s life” (Mercier & Doolittle, 2013, p.38). Also, according to Harris and Cale (2007), there is a relatively weak connection between youth fitness and health because children are still developing strength, endurance, and cardiovascular endurance. Social validity or user acceptance is another problem surrounding fitness testing, so
potential negative impacts (e.g., effect of low score on fitness engagement) of implementation should be considered (Harris & Cale, 2007; Mercier & Doolittle, 2013). The literature suggests that the use of fitness test scores along with BMI are inappropriate measures to determine teacher effectiveness in physical education, so used alone these are likely to be opposed by educators (Mercier & Doolittle, 2013). Inclusion in a portfolio can alleviate this pressure when the assessment is used as part of a more comprehensive, goal-oriented plan.

A final consideration for use of standardized fitness testing is expense. The cost of the Fitnessgram is approximately $600 per site license plus an annual subscription fee of approximately $150 (Human Kinetics, 2015). However, the Presidential Youth Fitness Program provides awards that allow schools to purchase the Fitnessgram license and pay the associated fees. Perhaps because of cost, some states (e.g., Connecticut) have opted to create their own fitness tests.

| Table 4-2. Fitnessgram measures of health-related fitness components. (Human Kinetics, 2015) |
|-----------------------------------------------|-----------------------------------------------|
| Component                     | What it Measures                                      |
| Cardiovascular Endurance      | Tests the ability of the cardiovascular system as well as the pulmonary system |
| Muscular strength             | Tests the amount of force that a muscle or muscle group is able to produce at a given time |
| Muscular endurance            | Tests the ability of a muscle or muscle group to continue contracting over a long period of time |
| Flexibility                   | Tests the movement possible around a joint           |
| Body Composition              | Tests the amount of lean tissue in comparison to the amount of adipose tissue |

(7) Checklists

Checklists provide a list of steps or features for students to complete a given task or learning objective. They are similar to rubrics in providing explicit outlines for student work,
but differ in that rubrics provide various levels of expected performance for grading. A checklist may be formative or summative, and is a useful tool for providing students with a detailed outline for a particular learning expectation or assignment (Rowlands, 2007).

Sallis and colleagues (Sallis, Strikmiller, Harsha, Feldman, Ehlinger, Stone, Williston, & Woods, 1996) collected students’ self-reported physical activity by using checklists (e.g., Physical Activity Checklist Interview, Self-Administered Physical Activity Checklist; SAPAC) containing various activities as well as the length and intensity of the activity (see Appendix 4B). They compared these checklists with objective measures of physical activity (i.e., accelerometer, heart rate monitor) and concluded that self-report checklists are a valid measure of students’ physical activity levels. The SAPAC is also reliable and valid in different languages when compared with objective measures (Gioxari, Kavouras, Tambalis, Maraki, Kollia, & Sidossis, 2011).

There are various benefits and drawbacks associated with the use of checklists in an academic setting. For instance, compared with objective measures, checklists require less time and monetary resources (Sallis et al., 1996), and are a useful way to conceptualize and distribute tangible guidelines or criteria to promote student success. Because of their ease of administration, checklists can be administered pre- and post-course to show changes over time (Angelo & Cross, 1993). Checklists also help students to be more systematic when attempting a task and during problem solving (Schraw, 1998) and help students become self-regulated learners (Zimmerman, Bonner, & Kovach, 1996). When completing difficult tasks, checklists facilitate learning and increased performance by specifying an outline or specific steps to complete (Delclos & Harrington, 1991; King, 1991).
For younger students, self-reporting via checklists may be difficult. For example, preadolescent children have trouble accurately recalling physical activity levels, especially after a longer period of time (Sallis, 1991). Checklists also do not address quality of work or activities, only the absence or presence of the checklist content (Rowlands, 2007). Dawson (2009) pointed out that peer-reviews guided by checklists may also limit authentic discussion, which may be an important component in visual or performing arts when discussing interpretations or reactions to assignments.

(8) Rubrics

Rubrics are becoming one of the most popular methods of assessment for performance of students in schools (Moskal, 2000). Rubrics describe specific criteria of levels of performance to assist in judging the quality of performance on a given task. According to Perlman (2003), a rubric is composed of one or more categories which are rated on a specified scale to reflect performance on a given dimension. Indicators or descriptors that demonstrate what is measured are included. Rubrics can be divided into at least two categories: holistic and analytic (Mertler, 2001), with most researchers choosing to divide them into task-specific and general categories, as well (Arter, 2000; Moskal, 2000; Perlman, 2003).

Types of Rubrics

Holistic rubrics (see Appendix 4C) assess along a single descriptive scale that provides a picture of overall quality of performance, while analytic rubrics (see Appendix 4D) allow for separate evaluation along different dimensions (Moskal, 2000; Perlman, 2003). Task-specific rubrics are tailored to specific assignments while general rubrics can be adapted to a variety of
assignments (Arter, 2000). Task-specific rubrics have the benefit of improved consistency of scoring and are considered better for assessing knowledge. General rubrics, on the other hand, are better for assessing more broad skills and products, and aid in helping students understand what constitutes quality on a greater range of tasks. Holistic rubrics are easier to use for large-scale assessment because they are less time-consuming, but are considered to be less useful for instructional purposes because they only provide a general picture of performance rather than the more detailed picture provided by analytic rubrics (Bargainnier, 2003; Perlman, 2003).

However, for simple performances where a single dimension can describe the quality of performance they can be useful (Arter, 2000).

**Quality Rubrics**

The literature shows that quality rubrics possess clear criteria, rich descriptive language, consistent wording, and focus on positive attainment. They also differentiate between performance, product, and effort and are valid and reliable (Bargainnier, 2003; Tierney & Simon, 2004). This means a superior rubric could have specific and well-defined criteria for each level of performance, and these criteria should accurately represent what is being measured (Bargainnier, 2003). Language should be concrete so the students will easily understand expectations. Positive attainment refers to what they should perform, rather than what students should not do. Finally, a proper rubric will make clear the difference between effort (whether or not an attempt is made), product (successful completion of an attempt), and performance (the quality of the attempt).

To maintain continuity within the rubric the descriptors between each scale point should be consistent, meaning the changes in descriptors reflect a difference in quality for the fixed
criteria (rather than introducing new criteria; Tierney & Simon, 2004). An example of a problem with consistency of a rubric would be if the performance criteria vary from level to level. An ideal rubric might also include examples of student performance at each level of proficiency (Perlman, 2003). McTighe and Arter (2000) defined anchors as “the representative products or performances used to illustrate each point on a scoring scale” (p. 1). The use of anchors to assist raters in the use of rubrics is recommended (Moskal, 2003; Moskal & Leydens, 2000; North Virginia Community College, 2008). Anchors also help clarify any uncertainties in the rubric that might be open to interpretation for both students and their guardians (Moskal, 2003). If there are three levels of performance within a rubric, it has been suggested to provide two anchors per level as examples to illustrate each level of mastery (University of Hawaii at Mānoa, 2013).

According to Jonsson and Svingby (2007), a reliable rubric has benchmarks, allows for analytical scoring, is topic-specific, and may require training to improve inter-rater reliability. They wrote that of the various types of validity, the assessment literature tends to focus on criterion, content, and construct validity. The authors described Messick’s (1996) six aspects of validity: content, generalizability, external, structural, substantive, and consequential. The content validity aspect of rubrics is most frequently assessed via opinions of subject matter experts. Generalizability of rubrics has been examined via comparing student scores across grade levels, while external validity is more often reported on as correlations between the given rubrics and other measures that purport to measure the same construct. The structural aspect of validity refers to whether the rubric and domain structure match up well, and has been judged via factor analysis or by raters who appraise whether the rubric lines up with the underlying standards or guidelines. Substantive validity refers to “theoretical rationales for, and empirical
evidence of, consistency in responses that reflect the thinking processes used by experts in the field” (p. 137). The authors reported that this concept is not assessed as often because most rubrics focus more on final products or performances rather than processes. Finally, the consequential aspect of validity refers to the implications of the score interpretations of the rubric, intended and unintended. In their review, Jonsson and Svingby (2007) found few rubrics being evaluated for consequential validity, but reported one study where raters indicated the new rubric they assessed to possess more instructional potential when compared to another rubric, as an example.

**Adapting an Existing Rubric**

When choosing an existing rubric, Perlman (2003) suggested asking whether the rubric being considered is related to the outcome(s) being measured or whether it is measuring anything unrelated to the construct at hand, if the criteria align with current ideas of what constitutes excellence in the field, whether the rubric generalizes to various tasks, whether the descriptions of the categories or scales are thorough and clear, whether the basis for assigning scores for each scale point is well defined, whether the rubric is easy to understand and can be scored consistently among different raters, and whether the rubric is practical to use. In order to adapt an existing rubric, one can consider altering the wording of another rubric, including changing or omitting a scale on an analytic rubric. Dividing a holistic rubric into more scales or using scales from several rubrics, is an option as well.
Developing a Rubric

When it comes to developing a new rubric, rather than adapting a preexisting one, Perlman (2003) recommended coming to a consensus with colleagues on what dimensions (criteria) of the final product or performance that must be assessed (possibly based on national curriculum frameworks). Moskal and Leydens (2000) recommended making the purpose and objectives of the assessment clear from the beginning and then developing scoring criteria that address each objective. Some literature supports the use of student input for creating rubrics and getting input from students on ideas for descriptors (Gibbons & Robinson, 2005; Huffman, 1998; McCollister, 2000). Then, it is suggested that actual student work is examined and divided into categories based on overall quality: works exhibiting the best quality, worst quality, and works that are considered to be in the middle (Perlman, 2003). Colleagues should discuss what makes the best works good and these qualities should be included in the criteria brainstormed in the initial stage. At this point it should be decided whether an analytic or holistic rubric is best for collecting the information needed for evaluation (Moskal, 2000). When the criteria that are going to be evaluated overlap, it may be best to use a holistic rubric for scoring.

The list of criteria should be sorted into categories or scales, or a single holistic scale can be developed (Perlman, 2003). For holistic rubrics, it is suggested that detailed descriptions relating to superior quality performance or product and inferior quality performance or product be written and that each attribute be combined in one comprehensive summary, then levels of performance should be thoroughly described on a continuum of best to worst performance and include the collective attributes being assessed (Mertler, 2001). For analytic rubrics, detailed descriptions of superior quality performance or product and inferior quality product or
performance should be written, for every individual attribute, then levels of performance should be thoroughly described on a continuum of best to worst performance with individual descriptions for each attribute.

Wolf and Connelly (2007) recommended three to seven criteria (to keep them more readily memorable), three to six levels of performance (few for summative assessment, and more for formative assessment), and using different language for the levels of performance headings (depending on whether the goal is assessing growth or mastery). Definitions of each dimension should be written out and should illustrate exactly what the dimension covers (Perlman, 2003). The next step is to develop a scale for each dimension that represents the range of possible products or performance, select student works that best represent each scale point, and write out a description of the best, worst possible, and intermediate products or performances. Then the rubric should be pilot-tested on actual student works, to see if the rubric is practical for use, and revised if necessary.

Arter (2000) suggested evaluating rubrics after they have been created for content, clarity, practicality, and technical soundness. When looking at content of the rubric, creators of the rubric should consider whether we can answer why each characteristic of the rubric is indicative of quality performance and whether or not we can cite field-relevant references on what performance is. These creators should be able to describe what is not included in the rubric and know why it was not included. When considering the clarity of the rubric creators should consider whether two different raters would give the same rating on the same product or performance and whether or not each aspect of the rubric could be explained to a student. In order to assess the practicality of the rubric, creators of the rubric should ask themselves whether
or not the rubric provides the information needed to guide instructional decisions and track students’ progress, as well as whether the rubric might be useful for encouraging improvement and assessing desired outcomes. Finally, when reflecting on the technical soundness of the rubric, creators should consider whether the rubric can obtain at least 65% precise rater agreement and 98% on one aspect, whether the language in the rubric is unbiased and appeals to a variety of learning styles, and whether the wording of the rubric truly describes performance rather than simply making judgments.

**Rubrics by Hard-to-Measure Area**

*Music Education.* Much of the literature on rubrics in music education focuses on developing new rubrics or on the general usefulness of rubrics in music education classrooms rather than reviewing existing rubrics (Hickey, 1999; Leonhardt, 2005; Wesolowski, 2012; Whitcomb, 1999). Whitcomb (1999) recommended creating different rubrics depending on the grade level of the students and ensuring that the descriptors are age appropriate. Student input for creating rubrics is suggested and may facilitate understanding of what quality music performance is.

Deluca and Bolden (2014) made an important contribution to the literature on music assessment rubrics by making the distinction between discrete-component rubrics and integrated-component rubrics, as well as elucidating the usefulness of self-reflection rubrics (see Appendix 4E). Discrete-component rubrics “measure one individual component of music at a time and accurately distinguish performance at various levels” (p.72), while integrated-component rubrics “aim to address the common challenge of how to assess a holistic performance in which
components combine and support one another, rather than separately assessing discrete performance components” (p.73).

Discrete-component rubrics are useful for helping students understand and aim for a higher level of performance. For evaluative purposes, the descriptive nature of discrete-component rubrics is also likely to increase the reliability of scoring, however these rubrics have the disadvantage of making musical components seem unrelated. Integrated rubrics focus on the relationship between musical components and are useful for teaching students how various components integrate to form a quality performance and if they are created using appropriately descriptive criteria, succeed both in enhancing reliability in assessment of student performance across levels and in providing feedback for improvement. The researchers suggest using comment boxes to clarify details with rubrics that cannot be adequately represented through the rubric alone and using student input when creating the rubric. Self-reflection rubrics for music are useful for developing students’ self-assessment skills. It is advised to use a rubric with performance levels leading from students being able to assess others’ music using performance criteria, to students assessing their own music and changing their performance accordingly when prompted by a teacher, to students assessing their own music without being prompted and progressing their performances.

Latimer, Bergee, and Cohen (2010) examined the reliability of a rubric created by the Kansas Music Educators Association to measure music performance in large group festivals. The rubric was found to possess moderate to moderately high reliability and was indicated as being useful to the raters who used it. The rubric was created to assess choir, orchestra, and band, included five levels of performance, and nine dimensions for assessing performance: tone,
intonation, expression, technique or diction (technique for band or orchestra and diction for chorus), rhythm, note accuracy, balance, blend, and other (which includes descriptors like posture and appearance).

Ciorba and Smith (2009) investigated the reliability and usefulness of a rubric for music performance assessment used at a Midwestern university and found the rubric to possess moderate to high inter-rater reliability. The rubric was created by four experienced music faculty members with between five and twenty years of experience teaching music. The rubric is on a one to five point scale, and divided into the categories of musical elements, command of instrument, and presentation which were found to be common to all vocal and instrumental areas. This rubric uses descriptors such as “Precise demonstration of musical elements is demonstrated” (p. 9), which does not closely align with the idea of using more objective descriptions of criteria for performance.

Wrigley and Emerson (2013) developed and validated a music performance scale for the five instrument families. The Performance Examination Rating Scale (PERS) was created with the input of subject matter experts and based off 15-17 music criteria for each instrument family. The piano rubric is divided into categories of technical mastery and control, sound quality, and convincing musical understanding. The levels of performance are labeled as needs attention, satisfactory, and excellent with a continuum of two sublevels of possible rating within the needs attention and excellent levels, and three sublevels within satisfactory. The researchers used structural equation modeling to investigate validity of the measures and evaluated the internal consistency as well. Based on their findings, they recommend against using generic assessment because of differences specific to the instrument families.
Visual Arts Education. The use of rubrics for assessment in visual arts classrooms has been recommended (Eshun & Osei-Poku, 2013; Huffman, 1998; McCollister, 2002). Students have found rubrics for art assessment beneficial for the learning process (Eshun & Osei-Poku, 2013). Connelly and Wolf (2002) reported that rubrics for art assessment support active learning, assist students in reaching the desired outcomes, and encourage stability in evaluations between faculty members. Huffman (1998) recommended creating rubrics for art assessment with the input of students and found that students’ involvement in rubric creation increased their sense of empowerment and commitment to the assessment process. McCollister (2002) also suggested student involvement in creating rubrics for art classroom assessment and acknowledged that students come into the art classroom with varying levels of skill and therefore it may be helpful to identify students’ entry point on a given rubric and encourage their progress.

Connelly and Wolf (2002) created a rubric for assessment of advanced painting for a midterm examination consisting of four categories of criteria and using a one to four point scale to evaluate each of the categories (Appendix 4d). The categories that were assessed are inventiveness, craftsmanship, productivity, and sketchbook. Inventiveness is described as taking risks, solving problems, and using a unique style. Craftsmanship is described as successfully using various materials and tools, presenting work professionally, and researching historical and cultural art examples. Productivity is described as creating a sufficient amount of art work, efficiently using classroom time, and seeking feedback from peers and the teacher. The sketchbook category is described as the student using a sketchbook to plan drawings, containing a set number of pages used, and evidence of studying famous artists. The rubrics categories have differing weights based on their importance, with each category being set at 25% of the grade, with the exception of the sketchbook category, which is given a 10% weight.
Physical Education. The use of rubrics for assessment of performance in physical education is supported (Birky, 2012; Dyson, Placek, Graber, Fisette, Rink, Zhu, & ... Park, 2011; Hensley, 1997; Schincariol & Radford, 1998). Students find the use of rubrics helpful for understanding what is expected of them and what they can improve on specifically (Birky, 2012). Hensley (1997) proposed using holistic means of assessment more often for assessing overall quality of game play of sports, and using analytic means of assessment for evaluating independent skills. Holistic assessment is a faster and more efficient means of assessing total performance in physical education classrooms.

Dyson et al. (2011) created rubrics for assessing kindergarten, second grade, and fifth grade in the physical education classroom. They assessed students’ demonstration of movement patterns and locomotor skills needed for various physical activities. The rubrics were created on the basis of one of the National Association for Sport and Physical Education standards for physical education and evidence for content validity was collected by consulting a team of subject matter experts on their opinions of the alignment of the assessments with the content standards. The rubrics were pilot tested and modified accordingly and item analyses revealed the rubrics to be well developed and (with some changes) ready for final calibration using a large national sample. One of the rubrics for measuring the standard “demonstrates competency in motor skills and movement patterns needed to perform a variety of physical activities” (p. 110), uses students’ ability to strike a ball 5 or more times consecutively with a short-handed paddle as the assessment task. The rubric uses five levels of performance and measures both successful completion of the task, and the quality of performing the task.
Multiple best-practices related to specific measures that can be used to assess student affect and performance were presented throughout this chapter. As indicated above, there are pros and cons for each of the measures described. As such, we used this information as well as the expertise of multiple stakeholders to guide our final recommendations about the types of measures suggested for process-portfolios described in Chapter Five.
Chapter Five: Compiling Measures into Portfolio Assessment

Portfolio Assessment and Associated Artifacts

**Definition and Types.** Portfolios can be diverse in their themes and presentation, but are typically defined as a collection of artifacts (pieces of work) gathered over a period of time as a result of collaboration between a teacher and student (Arter & Spandel, 1992; Mills, 2009; Winick, Avallone, & Crovo, 2008). As stated in the Phase I report (DiLoreto, Morganson, Arruda, Kessler, Biddle, O’Neill, & … Kass, 2014), the rationale for incorporating students’ input is that by giving them a share in the assessment process, they are more likely to engage. This increased engagement is likely to occur because stakeholder participation in performance assessment is shown to facilitate communication and yield greater goal alignment between parties (see DiLoreto et al., 2014). In the hard-to-measure content areas of education, portfolios are favored for their representative nature of both the student’s progress and the teacher’s influence. Phase I of the current project found that significant percentages of municipalities are currently using portfolios in these areas (47% in art, 40% in music, and 23% in physical education (DiLoreto et al., 2014).

Grounding the portfolio content requirements on federal or state standards helps to align common expectations and learning goals for student success, regardless of region, resources, or individual instructor. Additionally, state standards may provide educators with specific criteria for inclusion of student work (Belgrad, Burke, & Fogarty, 2008). Belgrad, Burke, and Fogarty (2008) suggested that the use of portfolio artifacts helps connect student work to curriculum goals, align it to the education standards, and enable assessment through the use of rubrics and checklists. In addition to providing concrete evidence that students are attaining the educational
goals, the artifacts help lend the student’s “voice” through print portfolios or through audio and video recordings as part of electronic portfolios.

Artifacts may take various forms including student projects, drawings, recordings, journals, responses to items on questionnaires, awards, and other activity logs. In practice, there are multiple ways to implement portfolio assessment. The teacher can serve as the decision-maker and decide on the type of portfolio, the learning objectives that will correspond with artifacts, and then assign coursework accordingly. However, artifacts may also be selected by other stakeholders (e.g., students, parents, administrators) in order to represent important concepts or skills in each subject area (Belgrad, Burke, & Fogarty, 2008; Rolheiser, Bower, & Stevahn, 2000). Regardless of who selects the content of the portfolio, it is important that the inclusion of artifacts follow point-to-point theory, specifying the content domain to be assessed is essential to establishing validity (Asher & Sciarrino, 1974). In other words, the practical implication is that the artifacts chosen to be included in the portfolio should be directly representative of the desired learning outcomes.

In regards to standards-based application, Lund and Kirk (2010) framed the number and type of artifacts as being either representative (spans one unit), cumulative (several units), one-dimensional (one standard), or multi-dimensional (involving two or more standards) depending on the number of artifacts corresponding with single versus multiple standard(s). For assessing one or more school year’s worth of learning, either a product or process portfolio can work as one or multidimensional. In order to gain a clear assessment of proficiency, a one-dimensional approach of targeting each standard individually may be ideal. Similarly, a process portfolio could be useful in training elementary aged students, while a product portfolio would be more
adept at preparing middle and high school students (who typically have a skill specialization in these areas) for the professional world.

Artifacts will differ based on the developmental level of the student (Belgrad, Burke, & Fogarty, 2008). With guidance and depending on age, the student can select an item of work from among several that fit the artifact description (this may be a more useful skill to develop in middle and high school). Lund and Kirk (2010) suggested designated class time for curating to teach and accommodate this selection task. Consistent with the model discussed in the next chapter, within each content area, artifacts should show students’ competence in cognitive, affective, and performance domains (see Table 5-1 and Appendix 4F; Maryland Department of Education, 2006; Lund & Kirk, 2010). Specifically, the affective component (e.g., self-reflection activities) provides insight into the metacognitive processes in otherwise performance-based tasks and the cognitive domain ensures retention of the underlying concepts of performance (Hughes, 2008). In Maryland, teachers are encouraged to use such artifacts in multiple ways to guide instruction, such as serving as a tool for communicating with parents (Maryland

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Cognitive Artifacts</th>
<th>Affective Artifacts</th>
<th>Performance Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual Arts</td>
<td>Comparison paper between an artist’s biography, work, and cultural/historical</td>
<td>Reflection paper on the ups and downs of the process of creating a sculpture</td>
<td>Photos of three-dimensional work with completed peer assessment rubric</td>
</tr>
<tr>
<td>Music</td>
<td>Theory test on notating volume in written music</td>
<td>Pre- and post-interest inventory of music genres to assess affective change from class</td>
<td>Sound recording of a live performance with completed teacher assessment rubric</td>
</tr>
<tr>
<td>Physical Education</td>
<td>Labeling worksheet of muscle groups involved in specific physical activities</td>
<td>Journal entry following physical activity to assess engagement and application</td>
<td>Video recording of student-created dance choreography with completed self-assessment rubric</td>
</tr>
</tbody>
</table>
The number of artifacts included in a portfolio is largely reliant on which purpose is chosen, but across types and content areas there seems to be a minimum of five and a maximum of 15 artifacts. Lund and Kirk (2010) suggested at least one artifact per standard for a physical education portfolio (e.g., five competencies would require at least five artifacts). In an art education study, Dorn (2006) used a total of eight artifacts, which allowed for early- and late-semester evaluations of four artifacts per evaluation. New York State Summer School of the Arts (n.d.), being a highly applied setting, requires 10-15 artifacts in a portfolio format as part of their admission process.

Various unifying themes can dictate the nature of portfolio content; the fine arts policy of Maryland’s Department of Education summarizes it most concisely in terms of product versus process portfolios. A product-focused collection consists of a student’s best work to show evidence of competency in a given skill. However, the process-based portfolios used in Maryland are designed to track development over time (Arts Education in Maryland Schools Alliance, n.d.; see Appendix 5A). A process portfolio naturally requires examples of deficits (e.g., early work) in addition to more successful attempts, which sets it apart from the more concise version. The product portfolio is what a professional in the field would use, which may be appropriate for career development in the upper grade levels (Chappuis, Stiggins, Chappuis, & Arter, 2012; Taylor, 1994).

Related to the issue of process portfolios, Chappuis (2009) recommended that students show “before” and “after” pictures of their learning, which fits the typical growth portfolio criteria of having at least two data points from which to assess progress (p. 157). Ideally, this is
done with early and later pieces of work to show representations of pre- and post-learning (Mueller, 2014). This approach allows portfolio assessment to be structured around specific learning objectives by having samples of the student’s skill set before and after the content has been taught. Students, teachers, and parents should be able to compare initial work to the assignments completed as a result of instruction and practice. Individual development is the focus, which is considered fair assessment (Silveira, 2013).

Limiting the number of required artifacts may also support standardized evaluation, storage logistics, and help alleviate teacher anxiety regarding portfolio implementation (Dotson & Henderson, 2009). A straightforward format such as beginning and end-of-year artifact collection keeps the product organized. This way, the same rubric can be used to measure different students, whereas a variable or large number of artifacts could lead to grading difficulties and an increased workload for all involved as they strive to compare different products. Additionally, the early artifacts can serve as communication tools by alerting teachers to performance discrepancies and providing evidence of the student’s starting point for when modified interventions are needed.

**Rating and Evaluation**

Similar to issues of design, portfolio scoring should also be related to the purpose of assessment. The product versus process model also factors into the final evaluation of the portfolio, as it would change the criteria from seeking evidence of mastery to seeking evidence of improvement. For example, if the purpose of the portfolio is assessing a student’s progress, then two artifacts, from similar assignments but at different time points, should be collected and graded (Herman, Gearhart, & Baker, 1993). An example of student involvement in artifact
selection may be working with the teacher to translate state-standards into “best work” artifacts to showcase the student’s achievement of these standards. The wording of the rubric used to assess the work would also be impacted by the purpose (product vs. proficiency; Lund & Kirk, 2010). Regardless of the end goal, Mills (2009) denoted the importance of communicating expectations and guidelines at the beginning of and throughout the school year. Additionally, Wineburg (1993) recommended the use of interviews with the examinee to discuss components of their work with evaluators. For high stakes assessment by the state, some state departments require specific artifacts and provide rubrics for evaluation, which increases validity and reliability of portfolio assessment (Belgrad, Burke, & Fogarty, 2008).

The individualization benefits of portfolios could be challenged by standardization needs, but the design and evaluation methods can be streamlined to help resolve this problem. One such approach is for the state to determine (very) specific learning objectives, for which the teachers assign work that results in artifacts (Arter & Spandel, 1992; Chappuis et al., 2012; Paulson, Paulson, & Meyer, 1991; Taylor, 1994; Valencia, 1990; Valencia & Calfee, 1991). Using this completed work, the teacher can create a menu of potential artifacts from which the students can choose (MSDOE, 2006). This method naturally lends itself to teacher-graded artifacts, but the final evaluation is a more complex issue.

There are three main concerns of reliability and validity when scoring portfolio artifacts: consensus, consistency, and measurement estimates (Stemler, 2004). Rubrics help to align artifact grading across raters, ensure consistency in grading, and reduce errors in ratings by making the expectations and criteria explicit (Jonsson & Svingby, 2007). Literature on rubric rating and inter-rater reliability suggests that having three raters using a rubrics increased
reliability over having two raters (Johnson, McDaniel, & Willeke, 2000); however use of two raters to establish inter-rater reliability is popular (Saxton, Belanger, & Becker, 2012; Stemler, 2004). The use of two raters is considered to be efficient for producing fair inter-rater reliability levels (Jonsson & Svingby, 2007). For a review of rubrics, refer back to Chapter Four.

Training sessions also increase the validity and reliability of rating work samples by identifying common errors in ratings (e.g., halo, leniency, central tendency; Bracken, Timmreck, & Church, 2001; Jonsson & Svingby, 2007; Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999). In an educational setting, norming raters or calibration processes have been used to score student work consistently, including the use of anchor documents to standardize ratings. Holmes and Oakleaf (2013) recommended rubric norming to improve reliability among multiple raters. Rubric norming is a complex process used to increase agreement when using rubrics for assessment. A facilitator who is familiar with the rubric at hand, has looked at all of the student work samples, and has background information on the rater team, is necessary to make the process efficient. Facilitators explain their thinking behind the scoring of three to five examples that ideally represent the range of student work (good, poor, average). The facilitator works as a role model and walks the team through his or her own use of the rubric on a sample of student work, but also understands the value of rater consensus. Next, the facilitator guides the raters through assessing some preselected student works (representative of the range of student work) one at a time. The raters all evaluate the sample of one to three selected works individually, and then the facilitator asks the group aloud what score they gave on a specific dimension of the rubric. The group members raise their hands if they scored the example a one, two, three, etc. This process allows the facilitator to see where the group is in agreement and what areas of the rubric will require more discussion. The facilitator leads a discussion of what areas of the rubric
have not reached much agreement and raters explain the rationale behind their scores. Then the facilitator helps the raters move towards agreement. The raters score a new set of example works, applying any new understanding gained, and ask any questions they may have to the facilitator. The facilitator repeats the steps of discussing scores assigned per example and per dimension, acknowledging dimensions on which there is agreement, and discussing the rationale behind disparate ratings with the goal of coming to an agreement. This process is repeated on areas of disagreement until raters reach consensus.

The norming process is often used when rubrics are being used, including at the state level. For example, Rhode Island conducted a 2-3 hour training session in which teachers examine student works, clarify questions, score the work, share the scores, and discuss scoring differences (Rhode Island Department of Education & the National Center for the Improvement of Educational Assessment, Inc, n.d.). The Advanced Placement Examination in Studio Art is also graded by a trained panel of judges using scoring rubrics, and a similar method was used in a study on student portfolio assessment, which resulted in high reliability of ratings (Dorn, Madeja, & Sabol, 2004). Score estimates may be inflated or deflated based on a variety of factors. Mabe and West (1982) found that self-assessments had low validity when compared to objective ability scores as they may be inflated or more lenient (Campbell & Lee, 1988). Peer-assessments may differ based on their own high- or low-performance status (Saavedra & Kwun, 1993), but they have the opportunity to observe other students’ performance and efforts without influencing the occurrence of these efforts as seen in teacher observations or expectations (Rosenthal, 1985). Multiple sources of feedback, aligned with rubrics and training, help to provide a reliable, valid assessment of student portfolios.
The formal method of assessment typically used for portfolios is a panel of judges who use a standardized rubric or checklist to provide ratings (National Assessment Governing Board, 2008). Informally, teachers can use the same criteria themselves. Castiglione (1996) pointed out that training evaluators about issues of consistency, fairness, and accuracy could be costly and time-consuming for large scale portfolio assessment. Other states have addressed these assessment issues by using Race to the Top funds to create regional committees of qualified personnel (USDOE, 2014). These teams are assigned to a number of schools to train faculty, supervise implementation, and collect data for new assessment practices.

Evaluation practices of both individual artifacts and overall contents appear to be widely distributed depending on the use and scope of the portfolio, but the options from specific to broad would be one rubric per artifact (related to the domain it addresses), one to three rubrics per competency (to encompass one or more domains), or one comprehensive rubric that addresses each competency and domain. Michigan Portfolios (2011), a multi-district division of the National Writing Project that has partnered with the Portfolio Research Group in Michigan, uses one comprehensive rubric per grade level. In this rubric, 5-6 skill sets are assessed on a scale of 1-5 and the numerical ratings are recorded on an assessment data sheet. This way, data to inform this project can be collected in a uniform method despite the individual differences between products. The rubrics were created through collaboration between Iowa and Michigan school districts and made available to the public online, where educators and other stakeholders can access them. The researchers proposed that rubric-based portfolio assessment makes it possible to achieve consistent and authentic evaluation results in a creativity-based area (Michigan Portfolios, 2011).
Presentation & Storage

Electronic storage and submission would enable third party scoring, while hard copy portfolios could be filed in the classroom and evaluated by the teacher. A previous report for the current study recommended digital storage based on practicality in artistic content areas, and the performance evidence needed for physical education can be viewed the same way. For example, several researchers (e.g., Asmus, 1999; Goolsby, 1995; Kelly, 2001) noted that electronic formats may be desirable in music so teachers can listen to performances via audio recording. Prior decisions on artifact types will influence the need for actual recordings versus copies of programs, teacher ratings, and other records that indicate (rather than display) a performance.

Maryland’s Department of Education (2006) currently allows the teacher to choose the portfolio format, but is progressing towards a web-based version that can be accessed by parents and school administrators in addition to the teacher and student. Electronic and web-based formats are not necessarily one and the same; Montgomery and Wiley (2008) compiled a detailed list of available software with both online and offline options. KidPix is rated as the easiest to use, while Microsoft PowerPoint is one of the moderately difficult choices. The web-based programs are all rated as difficult to use. For the purposes of the state, electronic data collection could still be achieved through an online reporting tool regardless of the portfolio’s format.

Conclusion

In conclusion, the choice of purpose (product versus process) as applicable to the grade level will determine the number of artifacts and teacher involvement. Both types should be
based on clear learning objectives derived from the standards, and include evidence of cognitive, affective, and performance proficiency. Storage can be flexible as long as procedures for evaluating and reporting the data have been established. Finally, a standard evaluation is comprised of rubrics and/or checklists that use criteria directly from the standards. Scoring is typically done by teachers who have been trained for it, but could also be accomplished by panels of judges or regional committees hired by the state. Non web-based, electronic storage is recommended where feasible to accommodate grading by third parties, if that is the desired option.
Chapter Six: Aim Four Model Development

Models of Assessing Students in Music Education, Art Education, and Physical Education

The fourth aim of this project focused on identifying a model of assessment that may be used to evaluate overall student knowledge, performance, and affect in music, visual arts, and physical education. In order to determine the best model of assessment, we collected information from multiple sources: subject matter experts in each of the disciplines, a national survey of current practices in the fifty states plus the District of Columbia, a survey of Florida practitioners, and best-practices found in the literature. As we evaluated assessment models, we considered advantages and disadvantages to each component. Below, we discuss the methods that we used to develop our final recommendations and considerations for implementation. In this chapter, we consider model issues about what should be addressed (i.e., model components and weightings), whether growth and proficiency models should be used, and how scores should be reported.

Three Domains of the Assessment Model

For the purposes of this grant, our analyses focused on evaluating three domains of student learning: knowledge (cognitive-based), performance (skills-based), and affect (attitudinal-based). The three components are consistent with classic educational theory (1956), is consistent with current practice (e.g., National Association for Sport and Physical Education, 2009; New Hampshire Association for Health, Physical Education, Recreation, and Dance, 2007), and, as elaborated below, overlaps with frameworks used to build national standards. The three-component model is summarized below and is depicted in Figure 6-1.
Figure 6-1: Overall Model of Assessment

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

- **Domain 3: 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.

I. Assessing Student Knowledge

Student knowledge is typically assessed by some form of traditional assessment in order for students to demonstrate their understanding of the content. Typically the demonstration occurs via a written or oral quiz, test, or exam. Traditional assessment is often appropriate for assessing factual knowledge (Anderson, 1998). Although some of the literature includes essays
and short answer tests as traditional assessments, the majority of the literature limits traditional assessment to selected response, e.g., multiple choice and true/false (Berkeley University of California, n.d.; Dikli, 2003; Perlman, 2003). Some positive features of traditional testing include ease and speed of scoring and ease of comparison of results (Hambleton & Murphy, 1992). Traditional assessment focuses on the cognitive aspect of learning; however, it is not suitable for all types of assessment. For example it would be inappropriate for assessment of affect and performance (Anderson, 1998; Nazario, 2005). Traditional assessments, such as multiple choice tests, are considered to be objective methods of testing students’ knowledge. It is difficult, however, to assess higher-order thinking via selected response tests (Anderson, 1998; Anderson, 1999).

II. Assessing Student Performance

Student performance is often assessed by alternative and/or authentic assessments. Performance assessments are useful for assessing skills that are difficult to measure with traditional tests, such as music performance, locomotor skills, or creativity (Perlman, 2003). Performance assessments are comprised of a performance task and a method of scoring (Anderson, 1999). Alternative assessments allow for an examination of the process as well as the product, while traditional assessment may be limited in this aspect (Anderson, 1998). Alternative assessments include all assessments that are not in the traditional oral exam or paper and pencil test format. Examples of alternative assessments include self- and peer-assessments, projects, portfolios, and journal entries (Hambleton & Murphy, 1992). Alternative and authentic assessments are similar in nature; however, authentic assessments involve the exhibition of a skill or application of knowledge via performance in a realistic setting (Shelly, 2002). Several
specific examples of authentic performance-based assessments are: creating works of art in a visual arts classroom as part of a portfolio and being assessed with a rubric, being assessed on motor skills during game-play in physical education, and assessing singing performance in small groups during a music class (Chiodo, Frakes, Macleod, Pagel, Shuler, Thompson, & Watts, 1998; Dorn, Madeja, & Sabol, 2004; Nazario, 2005).

III. Assessing Student Affect

Student affect is often assessed by students evaluating themselves and then reflecting on their demonstration of knowledge or skills. Although these assessments are less common, there are multiple reasons for student self-evaluation and reflecting on their knowledge and skills. For example, students could assess how they perceived learning. Student affect can be assessed by students evaluating themselves and then reflecting on their demonstration of knowledge or skills (Savickienè, 2010; Senger & Kanthan, 2012). Other forms of affective assessment include questionnaires given before and after classes to examine changes in attitudes and values (Savickienè, 2010). Although these assessments are less common, there are multiple reasons for student self-evaluation and reflection on their knowledge and skills including increased confidence and motivation towards completion of tasks (McMillan & Hearn, 2008). Self-assessments and/or self-reflective writings are often included as part of portfolios regardless of the discipline (Kitts, 2003; Mills, 2009; Tomhave, 1999). In visual arts, written reflections and responses to open-ended questions on knowledge gained, the learning process, skill growth, and personal choices have been used as affective assessment techniques (Gitomer, Grosh, & Price, 1992). Affective assessment has also been used in music and physical education classrooms in the form of writing about music goals and writing about why certain physical performances were
chosen as part of a physical education portfolio within the physical education classroom (Kitts, 2003; Mills, 2009). Self-reflective writing is typically scored using rubrics (Chang & Chou, 2011).

Weight Recommendations Based on Models of Assessment

In order to determine the appropriate weight of each of the three domains, we synthesized information from multiple sources: current practitioners, subject matter experts, extant literature, national recommendations, existing standards and corresponding benchmarks, and the work of past RTTT consultants. Based on a synthesis of all of the before mentioned sources (see Appendix 4F – table of weights), we recommend music and visual art education be comprised of 40% performance, 35% knowledge, and 25% affective components across grade levels. Additionally, we recommend that physical education be comprised of 45% performance, 35% knowledge, and 20% affective components across grade levels. Additional details about our recommendations are available in Volume I of this final report. Below is a summary of the findings from the literature, national standards, and previous consultants.

Findings from National Organizations and from the Literature

*Federal.* The National Assessment of Educational Progress (NAEP) created the 2008 Arts Assessment Framework that it uses as a blueprint for its federal assessment model (Winick, Avallone, & Crovo, 2008). This framework stated that the knowledge and skills that make up the arts are intertwined and that one entails the other. However, the NAEP Arts Framework does not explicitly state which percentages of overall knowledge and/or skills are best assessed through knowledge-based, performance-based, or affective-based assessment. NAEP does,
however, link the “responding” process of the arts (both music and dance) with affective, cognitive, and physical behavior. The document describes creating as “generating original works of art” and performing/interpreting as “performing an existing work” (p. 9). Additionally, the framework states that “Students need to be able to place the arts in broader contexts to fully appreciate their significance” (p. 10). This concept of building appreciation implies the affective aspect of learning.

**The Arts.** NAEP makes recommendations for percentage of assessment between the three processes (creating, responding, and performing) for music and the two processes (creating and responding) for the arts based on specific grade levels. For example, NAEP recommends that 20-30% of music assessment in fourth grade should focus on creating, 40-50% should focus on performing, and 25-35% should focus on responding. In the twelfth grade, however, NAEP recommends that 20-30% of music assessment should focus on creating, 35-45% on performing, and 30-40% on responding.

NAEP makes similar recommendations for art assessment. Specifically, for art assessment in fourth, eighth, and twelfth grades, 50-70% should focus on creating and 30-50% should focus on responding. The percentages also reflect the division of time that should be spent in each of these processes within the respective discipline.

**Previous Florida Race to the Top Grant (RTTT) Project Findings**

**The Arts.** The Race to the Top Performing Fine Arts Assessment team recommended that 30% of students’ performing arts assessment should focus on the cognitive domain using
traditional assessments and 70% should focus on performance – combining creating and performing.

**Physical Education.** Research concerning physical education assessment has addressed model proportions. Specifically, Gallo (1999) completed a research study that focused on the appropriate weights recommended for each of the three domains. In her study, she formed an Action Research Study Team (ARST) that consisted of five physical education subject matter experts. She asked them to report the assessment techniques that they typically use in their classrooms. A total of five ARST members provided the following information: (1) Two members assessed the affective and cognitive domains at a 25% weight and motor (performance) at a 50% weight; (2) One member assessed the affective and cognitive domains at a 30% weight and motor domain at a 40% weight; (3) One member assessed the affective and motor domains at between 15 and 45%, and the cognitive domain at 30% weight; and (4) One member assessed the affective domain at 60 – 70% weight with cognitive and affective domains assessed at a 15 – 20% weight. Gallo (1999) reported that she would weigh the assessment at 25% each between the cognitive, affective, and motor domain and allow students to choose an area in which to assess the remaining 25%. The New Hampshire Association for Health, Physical Education, Recreation and Dance reported that it uses the following domains: affective-personal, cognitive, and psychomotor (New Hampshire State Department of Education, 2007). According to the New Hampshire State Department of Education (2007), students are assessed based on six guidelines, two of which are classified as being psychomotor, one both cognitive and psychomotor, two affective, and one cognitive.
Review of Existing Standards and Corresponding Benchmarks/Grade Level Expectations

One of the methods we used to determine the percentage contribution of the three domains (i.e., cognitive, performance, and affective) toward the overall assessment model included the following systematic approach. Our research team formed three groups to review the national and grade-level outcomes and State benchmarks for each of the hard to measure disciplines (i.e., music, visual art, physical education). Each group included three to four team members, with a discipline-specific subject matter expert assigned to each group. Each group was provided with a working definition of each domain (i.e., those listed as bullets earlier within this chapter). Group members independently reviewed the outcomes and benchmarks for every grade level within the discipline and sorted each of them into one of the three domains using the domain definitions as a guideline. Because several items included compound ideas, the reviewers could designate an item as fitting into more than one domain (e.g., both affective and cognitive). After group members individually sorted all of the items in their assigned discipline, they met as a group to compare results. If group members could not agree with the category or categories an item belongs to, then they discussed it until a majority decision could be reached. The item was then sorted into that category or categories. Once all three groups completed their sorting, the weights were determined by dividing the number of items in each of the three categories by the total number of sorted items. Because items that fell into more than one category were typically compound items, they were allowed to be counted more than once so that the percentage contribution of each domain would sum to 100%. The percentages were calculated as an overall model and by grade level (i.e., elementary, middle, and high school levels) such that different models could be employed based on differences across the three grade levels. See Appendix 6B for the final results of this coding process.
Growth versus Proficiency

In addition to focusing upon what should be measured (as addressed by the model components above), we considered issues of temporality in developing our model. Many features of a portfolio are dependent on whether a growth versus proficiency model is adopted. To ascertain the most viable method for these purposes, both types were examined. Growth models were determined as most fitting for assessing student achievement over a school year in addition to other variables, while proficiency models are only applicable to specific criteria related to student performance.

Proficiency Model. A proficiency (or status) model is considered a gauge of competency because it can measure how students perform at one point in time (i.e., a snapshot). It reports on performance by providing one data point per construct and compares the proficiency level with an established target. Students who do not meet the target criteria are not considered proficient. While competence rates are useful indicators of class standing, these models have been criticized as an inaccurate analysis of overall achievement of the student and school.

In contrast, a growth model accounts for the varying starting levels of individual students and focuses on improvement made from that point. It is a longitudinal method for measuring the amount of academic progress each student makes between a minimum of two points in time, as opposed to a single snapshot. To assess progress, this type of model requires the acquisition of baseline data prior to assessing growth. Sources of baseline data include state assessments, performance in previous classes, and work samples. With these data, teachers may construct student learning objectives, including strengths and weaknesses. The means of data collection
may be different at the different assessment points, and may be combined into an assessment portfolio.

**Growth Model.** Growth models are a means of tracking progress towards proficiency goals over time. This approach shows alignment with the fundamental goal of education, student learning. As stated by Secretary of Education Margaret Spellings, this approach ensures “greater flexibility in tracking individual students’ annual progress” and “provides states with more options for a nuanced accountability system, while adhering to the core principles of No Child Left Behind.” (USDOE, 2008, para. 1).

Limitations to using this model are reflected in the variability in measurement items and requirement of baseline data in addition to a growth target. Growth models are inherently more complex than a proficiency model, but more individualized. Also, there are different types of growth models and statistical analyses from which to choose. For example, Colorado has implemented a “projection” growth model that was designed to use growth percentiles from one year to predict whether a student will meet proficiency goals in forthcoming years (O’Malley, Murphy, McClarty, & McBride, 2011). Implementation may require further decision-making in regards to additional use of the data.

**Raw and Scaled Scores**

Raw scores are the unaltered, total number of points a student receives. Percentage-correct scores represent the number of points a student receives divided by the total number of points possible. Standardized testing entities use scaled scores, which are “obtained by statistically adjusting and converting raw scores onto a common scale to account for differences
in difficulty across different forms” (Tan & Michel, 2011, p. 3). Scaled scores are useful for standardizing scores to compare various assignments across difficulty levels. They also allow teachers to compare performance across various assignments. For instance, two separate paintings may be scored using different rubrics, resulting in incomparable raw scores. If these scores are converted to scaled scores, the teacher is able to compare performance scores over time on various assignments, and stakeholders are able to compare students’ standardized scores across schools. For example, the AP Exam in Studio Art Drawing requires student portfolios, and each artifact in the portfolio is rated by multiple evaluators on a 6-point scale. The combined scores are later transformed into a final 1-5 score (College Board, 2015).

The benefits of raw scores and percentage-correct scores are the ease of calculation and comprehension, but a drawback is the possibility of misinterpretation. Percentage-correct scores may only be used to compare students in the same grade level, but this limitation impacts grading and comparison at different grade levels. Scaled scores are useful for this purpose, but may be difficult to calculate. Scaled scores can be used to assess differences in performance among districts or groups of students, and year-to-year growth (Smarter Balanced Assessment Consortium, 2014). For example, difficulty in one subject area may be identified by comparing performance across grades and districts. If one grade level or district falls behind when comparing the scaled scores, a targeted intervention could be implemented for improvement at this grade level or district. Scaled scores are more difficult to calculate, but are commonly used in standardized testing (e.g., AP exams, SAT, GRE) and the scores are more accurately used and interpreted (Tan & Michel, 2011). Damiani (n.d.) pointed out that portfolio usefulness is limited by the interpretation of results by stakeholders. Scaled scores address this limitation, making the
interpretation of student’s portfolio scores more straightforward for educational decisions across students, grades, and schools.

There are two ways to interpret students’ scaled scores, criterion-referenced and norm-referenced. Criterion-referenced interpretations represent a level of mastery of the material, scaled to represent the student’s level of competence of the particular material. Norm-referenced interpretations are scaled scores used to compare the student’s score to a norm group, or similar sample of students. This group may be at the national, local, or subgroup level (i.e., private, specialized, or large city schools). The choice between these interpretations should be based on the purpose of the comparison: criterion-referenced to determine a student’s success in meeting curricular aims, norm-referenced to rank order students compared to their peers. In both cases, cut scores are often used to place students into pass/fail categories based on acceptable scores as determined by a team of experts. Statistically determining appropriate cut scores each year based on previous test results is imperative to keep pass/fail decisions accurate and fair.

Summary

As evidenced throughout this report, our team completed a methodologically-sound, systematic approach to investigating the best way to assess student knowledge, performance, and affect. As such, it’s important to consider the limitations that were addressed throughout the report and in the recommendations section/Volume 1. Specifically, due to the nature of the three disciplines and our recommendations; it’s likely that the implementation of the recommendations will take a tremendous amount of time and resources. Although we provide the general guidelines for implementation, it’s likely that the districts will be faced with legitimate challenges making it extremely difficult to implement the recommendations. As such, it is
important that the FLDOE realize both the strengths and the limitations provided throughout Phase II of this current project.