2017

Criteria-Referenced Formative Assessment in the Arts

Fei Chen
University of North Carolina at Chapel Hill

Angela Lui
University at Albany, State University of New York, alui@albany.edu

Heidi Andrade
University at Albany, State University of New York, handrade@albany.edu

Christopher Valle
3M Health Information Systems

Hirah Mir
University at Albany, State University of New York, hmir@albany.edu

Recommended Citation

Chen, Fei; Lui, Angela; Andrade, Heidi; Valle, Christopher; and Mir, Hirah, "Criteria-Referenced Formative Assessment in the Arts" (2017). *Educational & Counseling Psychology Faculty Scholarship*. 10.
https://scholarsarchive.library.albany.edu/edpsych_fac_scholar/10

Criteria-Referenced Formative Assessment in the Arts

Fei Chen, Angela Lui, Heidi Andrade, Christopher Valle, and Hirah Mir

University at Albany—State University of New York

In press, *Educational Assessment, Evaluation & Accountability*

Abstract

The purpose of this study was to examine the effect of criteria-referenced formative assessment on achievement in the arts. Seventy-five schools in New York City were assigned to either the treatment or control condition. The treatment involved 3,195 elementary, middle or high school students instructed by 43 music, visual arts, theater, or dance teachers. The teachers were involved in a professional development program focused on formative assessment practices, particularly criteria-referenced peer and self-assessment. The control group consisted of 2,445 students in classes instructed by 32 teachers who did not receive the professional development. Discipline-specific, performance-based pre- and post-measures were used to evaluate student learning. Fidelity of implementation was examined before the analysis of the treatment effect. Propensity score matching analysis was used to examine group differences in performance on the post-assessment. Results based on a sample of 611 matched pairs of students showed that, overall, criteria-referenced formative assessment had a statistically significant, positive effect (d =.26) on students' arts achievement.

Keywords: arts education, criteria, formative assessment, peer assessment, rubric, self-assessment.

Authors' Note: We wish to thank Dr. Susanne Harnett and Tara Mastrorilli of Metis Associates for sharing their data and wisdom, Drs. Jason Bryer and Robert Pruzek for guidance on using propensity score analysis, and Dr. Kara Hogan for help with the rating of implementation logs. This research was supported in part by a faculty research award to Dr. Heidi Andrade from the University at Albany, SUNY.

An Empirical Investigation of Criteria-Referenced Formative Assessment in the Arts

This study investigated the effects of criteria-referenced formative assessment on students' achievement in the arts (dance, music, theater, and visual arts), using a pre-post randomized block design. Elementary, middle, and high school teachers in the treatment condition received professional development emphasizing the use of technology and formative assessment practices. Formative assessment was conceptualized in terms of students as both producers and users of assessment information (Andrade, 2010; Black & Wiliam, 2009), and operationalized as teacher feedback, peer assessment, and self-assessment according to clearly articulated task criteria. Arts achievement was measured by art-specific performance tasks.

We had one research question: Is there a difference in arts achievement between students whose teachers engaged them in criteria-referenced formative assessment and those who did not? We hypothesized that criteria-referenced formative assessment would increase students' achievement in the arts even when controlling for pre-treatment achievement and key demographic characteristics.

Formative Classroom Assessment

Formative classroom assessment is the practice of using evidence of student learning and achievement to make adjustments to instruction and learning strategies in order to better meet students' needs (Wiliam, 2010). The operationalization of formative assessment as criteria-referenced feedback drew on Sadler (1989) and Hattie and Timperley (2007), who characterized formative feedback in terms of three questions to be asked by teachers and students: "Where are we going? Where are we now? Where to next?" According to Wiliam and Thompson's (2007) feedback systems model (Figure 1), formative assessment that promotes learning involves uncovering a gap between the current and desired states of learning, and determining ways to close the gap. Wiliam and Thompson's model acknowledges the key role students can have during assessment. Through assessment, students can help each other identify areas in need of improvement, and serve as resources for moving their performance closer to the criteria for success. We added detail to Wiliam and Thompson's (2007) model in order to emphasize the potential of students as sources of feedback in the assessment process (Valle, 2015). In our version of the model (Figure 2), students serve as instructional resources for themselves and one another, actively engage in assessing the quality of their work in relation to the success criteria, and provide detailed and specific feedback to close gaps in learning.

In order to determine if there are gaps in their understanding, students need to know what qualifies as good learning. Therefore, an essential step in the feedback system model is communicating learning intentions and criteria for success, and ensuring that students understand them. Furthermore, the quality of students' learning can only be assessed when they engage in a task designed to demonstrate their learning. Thus, formative assessment involves communicating explicit and clear task criteria that students reference throughout the entire assessment process.

This study therefore emphasizes three key strategies of formative assessment: (1) clarifying and sharing learning goals and criteria for success; (2) providing feedback that moves learners forward; and (3) promoting students' roles as instructional resources for their peers and themselves. We refer to this conceptualization of formative assessment as *criteria-referenced formative assessment*.

Criteria-Referenced Formative Assessment

The focus on criteria-referenced assessment is important because it addresses the question, "Where are we going?" Success criteria describe the qualities of excellent student work on a particular assignment and can be communicated to students in a variety of ways. Worked examples, which typically consist of a sample problem and the appropriate steps to its solution, imply success criteria. Hattie's (2009) meta-analysis resulted in an overall effect size of worked examples of *d* = 0.52.

Direct expressions of the success criteria include rubrics and checklists. Andrade and colleagues (Andrade, Du, & Mycek, 2010; Andrade, Du, & Wang, 2008) used rubrics to communicate success criteria to the elementary and middle school students in studies of self-assessment of writing. Students read a model essay, discussed its qualities, and generated a list of criteria that were then included in the rubric they used to self-assess drafts of their own essays to inform revision. Scores for the treatment group's essays were practically and statistically higher than those of the comparison group, with effect sizes of *d* = .87 and .66.

Ross and Starling (2008) also ensured that the grade nine geography students in their study understood and could apply the criteria for assessment to their own work. Before asking students to self-assess their projects, students were involved in defining assessment criteria by co-constructing a rubric, and learned to apply the criteria through teacher modeling. After controlling for the effects of pre-test self-efficacy, students in the self-assessment group scored higher than students in the control group on all of the achievement measures, which included a Global Information System map, a report, and an exam. The self-assessment treatment accounted for 22% of the variance across achievement measures (Ross & Starling, 2008).

It is important to note that the studies by Andrade et al. (2009; 2010) and Ross and Starling (2008) involved self-assessment as well as transparent success criteria. A search of the literature revealed only one study conducted in a K-12 context that examined the effect of success criteria alone: it was a study Andrade (2001) conducted on the effects of simply providing a rubric to eighth grade writers before they began to write. Of the three essays students wrote for the study, only one resulted in significant differences between the treatment and comparison groups. In a higher education context, however, Lipnevich, McCallen, Miles and Smith (2014) found that providing 100 undergraduates with rubrics, exemplars, or both was associated with significant improvements in student performance, with rubrics edging out the exemplars-only and rubrics+exemplars conditions in terms of effect size (rubric only Cohen's *d* = 1.54; exemplars only Cohen's *d* = 1.04; rubrics + exemplars Cohen's *d* = 1.04). The students in the Lipnevich et al. study reported that they used the rubric and/or exemplar to guide revision, which implies they engaged in an informal self-assessment of their work. Given the results of these studies, it seems reasonable to assume that sharing success criteria with students should be part of a comprehensive process of actively engaging them in assessment through self- and/or peer assessment.

Peer and Self-Assessment

Peer and self-assessment are formative assessment techniques that have shown promise in terms of student learning and performance (Andrade, 2010; Brown & Harris, 2013; Topping, 2013). Their effectiveness is likely due to the fact that, when carefully implemented, peer and self-assessment provide students with low- or no-stakes feedback on the quality of their work. That feedback addresses the questions, "Where are we now?" and "Where to next?"

Brown and Harris's review of the relationship between self-assessment and academic achievement in K-12 contexts included interventions across grade levels and in a variety of disciplines, including language arts, math, music, and high school qualification exams. The median effect size was between *d* = 0.4 and 0.45. Topping's (2013) review of peer assessment distinguished between elementary/middle and high school populations. Topping found relatively few studies that showed that peer assessment in elementary schools was related to increased achievement, but several studies conducted in high schools did indicate a relationship between peer assessment and achievement. Some of those studies suggested that peer feedback was effective for both the assessed and the assessors.

Formative Assessment in the Arts

Much of the research on formative assessment has taken place in core subject areas. This study examined its effects in arts education classrooms, including music, dance, theater, and visual arts. Although formal evaluation is anathema to many arts teachers (Colwell, 2004), key elements of formative assessment are inherent to artistic practice. For example, the rehearsal process, which is at the heart of music, dance, and theater, is an ongoing, formative assessment experience during which performers get feedback about their performances and revise accordingly. The critiques to which visual artists are often subjected also serve a feedback function.

The difference between traditional rehearsal and critique processes and the formative assessment processes employed in this study is related to the nature of students' involvement. Rather than simply taking direction, students were aware of the success criteria, actively participated in giving and receiving feedback intended to move themselves and each other toward their goals, and meaningfully engaged in rethinking and revising performances in the service of the goals. Given research on the association between formative assessment and achievement in core subjects (Bennett, 2011), it was predicted that criterion-referenced, formative assessment in the arts would be related to increases in student learning as measured by performance-based assessments of knowledge and skill.

Research on assessment in the arts has relied largely on anecdotal evidence and individual case studies. For example, Englebright and Mahoney (2012) presented anecdotes on implications for dance instruction when a performance assessment was implemented into a state education system. Harding (2012) examined formative assessment practices in dance education but focused on one classroom. There is a dearth of research on formative assessment in the arts that uses more methodologically rigorous designs.

Current Study

Reviews of research suggest that, when implemented well, formative assessment can promote student learning (Bennett, 2011). This claim is based primarily on studies conducted with relatively small sample sizes. This study contributes to the literature by investigating the effects of formative assessment on a much larger scale. Secondly, this study focused on formative assessment in the arts, an area in need of more large-scale empirical attention (National Endowment for the Arts, 2012). The study was part of a project called *Arts Achieve*. Funded by a U.S. Department of Education Investing in Innovation (i3) grant and an Arts in Education Model Development and Dissemination (AEMDD) grant, the project began in 2010.

Arts Achieve was evaluated by colleagues at Metis Associates, who designed the study and provided access to de-identified data on the schools (e.g., district borough number, school level, disciplines) and student demographics (gender, ethnicity, special education status, socio-economic status as measured by the free and reduced lunch indicator, scores on the New York State standardized mathematics and English/Language Arts achievement tests, average daily attendance, and whether or not students had received yearlong instruction in the art form). The data for this study came from the first year of the *Arts Achieve* data collection (Fall 2011 to Spring 2012).

Using the feedback systems model in Figure 2, criteria-referenced formative assessment was operationalized in terms of criteria-referenced feedback, and self- and peer assessment.

Consistent with the literature, criteria-referenced self- and peer assessment in this study involved the use of rubrics, checklists, and other tools that communicated learning goals and criteria to students.

Methods

Participants

Seventy-five New York City schools spanning all five boroughs and 36 districts were randomly selected from a pool of high needs schools to participate in this study. Schools were randomly assigned to the treatment or control condition by art discipline and school level. Students were not randomly assigned within schools. Archival demographic data from the New York City Department of Education were collected, including gender, ethnicity, math achievement, writing achievement, State English/Language Arts (ELA) test score, special education status, socio-economic status as measured by the free and reduced lunch indicator, average daily attendance, and whether or not the child had received yearlong instruction in the art form.

The total sample included 5,640 dance, music, theater, and visual arts students (control=2,445; treatment=3,195). Because a treatment condition with high fidelity is essential to examining the true effects of criteria-referenced formative assessment (CRFA), only those students whose teacher implemented the treatment with high fidelity were included in this study. After omitting students in the low-fidelity group, imputing missing values on several of the pre-test covariates, and dropping cases without a post-test score, the final sample was comprised of 2,219 students (control=1,608; treatment=611). Descriptive statistics for the control and treatment groups from the sample used for analysis are presented in Tables 1 and 2.

Instruments

Benchmark Arts Assessment. Pre- and post-assessments of achievement in the arts were developed for each of the four art forms. The Benchmark Arts Assessments (BAA) were created by Curriculum and Assessment Development teams comprised of educators, art teachers, and partners from several institutions, including the New York City Department of Education's Office of Arts and Special Projects, and Office of Tests and Measurement. The assessments were developed to authentically measure students' conceptual understanding, literacy, application of knowledge, and analytical and performance skills relevant to each art form. The majority of the tasks on the BAA were performance-based (e.g., choreographing and performing a dance, composing a short piece of music, creating a collage, writing and acting out a play script). The assessments also included multiple choice, short response, essay, and fill in the blank items. All tasks and scoring rubrics were aligned with the NYCDOE *Blueprints for Teaching and Learning in the Arts* and the Common Core State Standards in English Language Arts. Sample BAA performance tasks can be found in Appendices A-D.

Field trials of the Benchmark Arts Assessment were conducted in Spring 2011 on a pilot sample of New York City schools similar to the target study sample. Data from the pilot study were used to gather information about the psychometric properties of the assessment. Content experts in each art domain reviewed the assessments and found them to have face validity. The pilot data was used to compute internal consistencies of each form of the assessment: dance, α = .825 (15 items); music α = .888 (23 items); theater/acting, α = .868 (9 items), theater/musical theater, α = .825 (7 items); theater/playwriting, α = .891 (7 items); and visual arts α = .876 (25 items) (Metis Associates, 2015). In the summer of 2011, the assessments were revised, finalized, and prepared for administration during the 2011-2012 academic year.

Students received the pre- and post-assessment of the BAA in Fall 2011 and Spring 2012, respectively. Because several revisions were made to the BAA after the initial reliability analysis, we recalculated both the inter-rater reliabilities and internal consistencies using the Fall 2011 and Spring 2012 data. Inter-rater reliability for the Fall 2011 assessment administration ranged from .33 to 1.00; and from .11 to 1.00 for Spring 2012. We also calculated internal consistencies for each form of the assessment by school level. As shown in Table 3, the internal consistencies for Fall 2011 and Spring 2012 are generally acceptable, except for elementary and middle school visual arts and high school dance items from Fall 2011, which had low levels of reliability.

A close examination of the visual arts and dance assessments with low internal consistencies indicated that each of the dance and visual arts assessments, across all disciplines, were multidimensional, targeting an authentic variety of knowledge and skills. Since internal consistency assumes unidimensionality (AERA, APA, & NCME, 2014), items that tap multiple constructs are likely to produce low internal consistencies. Because we are most interested in the students' overall arts achievement, rather than achievement in terms of specific skills and knowledge, we included these assessments in our analyses.

Implementation Logs. Implementation logs were used to document the use of criteria-referenced formative assessment practices by teachers in the treatment condition. In their logs, teachers described how teacher feedback, peer assessment, self-assessment, rubrics, checklists, and other practices (e.g., technology use) were used throughout the learning process. A fidelity of implementation variable was created to assess the extent to which the delivery of criteria-referenced formative assessment matched the program's goals. Two researchers analyzed the implementation logs to identify treatment teachers who explicitly reported the use of: 1) rubrics, checklists, or other tools to share criteria with students, 2) teacher, peer, and/or self-assessment to judge and generate feedback about the quality of students' works-in-progress, and 3) opportunities for revision during which students could deepen their learning and improve their work.

If a teacher met a threshold of evidence of these activities in his or her log, that teacher was coded with a "2," indicative of high fidelity of implementation. All other teachers in the treatment group were coded with a "1," meaning that they received CRFA training but did not show evidence of high fidelity of implementation during the first year of the project. The raters discussed the codes until there was 100% agreement on the coding of teachers as either high or low fidelity. All teachers in the control group, who carried out business as usual in their classrooms, were assigned "0" as their fidelity of implementation code.

Procedure

Students in the treatment and control conditions were administered the Benchmark Arts Assessment at the beginning and end of the 2011-2012 school year. As previously described, students in the treatment condition received instruction from teachers trained in CRFA and who delivered the treatment with high fidelity. Teaching to the test was not a concern given that the teachers trained in CRFA neither had information about the content of the BAA tests, nor were they present during the administration or scoring. Students in the control group received business-as-usual instruction, i.e., instruction from teachers who did not receive formal training on criteria-referenced formative assessment.

Data Preparation and Analyses

Missing Data Imputation. Missing data was imputed using the multiple imputation procedure in the MICE *R* package (Buuren & Groothuis-Oudshoorn, 2011). In order to avoid bias in propensity score estimation attributed to data imputation, missingness on outcome variables were not imputed: Cases with missing values on the outcome variables were excluded from further analysis. A comparison of the equivalence of cases removed from analysis and the final sample showed that, on average, the students in the final sample had higher prior achievement in math, writing, and ELA, higher average daily attendance (ADA), and higher pre-test BAA scores than those excluded from the sample due to missingness on the outcome variables. This finding constitutes a limitation of the study. The final sample for analysis consists of a total of 2,219 students: 611 in treatment and 1,608 in control. Descriptive statistics for the control and treatment groups from the sample used for analysis are presented in Tables 1 and 2.

Propensity Score Matching. Propensity score matching (PSM) with *R* (R Core Team, 2013) was conducted in two phases. In phase one, logistic regression modeling was conducted to estimate the propensity of individual students being assigned to the treatment condition. Covariate balances were examined after propensity score modeling. In phase two, students in the treatment group were matched with their counterparts with similar propensity scores in the control group in order to achieve a relatively unbiased estimate of the effect size of treatment on the outcome variable (performance on the 2012 BAA post-assessment). The binary treatment variable was treatment or control group.

Twelve variables theoretically associated with the outcome variable were selected as covariates in the propensity score model. Of the twelve variables, five are continuous: (1) performance on the 2011 BAA pre-assessment in the arts, (2) NYS 2011 test of English Language Arts score, (3) NYS 2011 test of mathematics score, (4) pre-assessment writing skills, and (5) average daily attendance. The remaining seven are categorical variables: (6) discipline,

(7) school level, (8) English Language Learner status, (9) special education status, (10) socio-economic status (free or reduced lunch), (11) ethnicity (White/Minority), and (12) gender.

A logistic regression with the covariates was used to model the probability of students being assigned to the treatment group. Using the propensity score model, observations from the two conditions were matched one-to-one with replacement. Partial exact matching was used to match students from the treatment condition with those in the control, while exact matches were assigned for discipline and school level. Within each subgroup by discipline and school level, nearest neighbor matching was used to match pairs of treated and control students based on the estimated propensity scores. The propensity score matching was done using the `Matchby` function [Matching] in R (Diamond & Sekhon, 2005).

Results

A total of 611 matched pairs of observations were obtained after propensity score matching. See Table 4 for a summary of distribution of pairs of students by discipline and educational level after matching. Student demographic information of the matched sample is summarized in Table 5. No high school music or middle school theater arts teachers were identified as high fidelity of treatment, so no student samples from these two subgroups were included in the estimate of the treatment effect.

Effect sizes of covariates between treatment and control groups before and after propensity score matching were calculated to examine the balance of the covariates after propensity score modeling. As shown in Figure 3, the initial balance of most covariates was acceptable before matching (*i.e.*, effect size of most covariates is below .20). After exact matching of the treatment and control groups by discipline and grade level, as well as nearest neighbor matching by propensity score, the balance of most covariates was further improved or remained similar to the initial effect size (see Figure 3).

Examination of the balance of covariates indicated that Fall 2011 performance score was the only covariate that had larger between-group effect sizes after matching than before matching. This is mainly due to the limited number of control group students to be matched with treatment group students in some subgroups after exact matching on discipline and grade level. Nonetheless, sufficient multivariate balance was achieved using the current propensity score model: The matched pairs in the treatment and control groups are similar and comparable in terms of the covariates included in the model, using the criterion of between-group effect size of less than 0.25 (Harder, Stuart, & Anthony, 2010). In sum, the propensity score matching model showed adequate balance and was employed to preprocess the data for estimating the treatment effect.

Using the propensity score matched sample, the control and treatment students' performance on the 2012 BAA post-assessment were compared. The overall average treatment effect on the treated (ATT) was *d*=.26 (95% CI = [.15, .37]), which was statistically significant (*t* (610) = 5.10, *p* =.00). The small effect size favored students in the treatment group.

Discussion

This study sought to gather evidence of the causal effect of criteria-referenced formative assessment on students' performance in the arts, and to do so at a large scale. Because the data was randomized at the school but not the student level, propensity score matching was used to adjust for the lack of complete randomization. Propensity score matching techniques also adjusted for the imbalance between groups on key covariates, and to strengthen causal inferences based on the findings. The use of propensity score methods in strengthening causal inferences of treatment effect is unprecedented in arts education and rare in research on classroom assessment in general.

Using propensity score matching analysis, this study examined the relationship between student learning and teachers' use of criteria-referenced, formative peer and self-assessment in the arts. Only an overall treatment effect was estimated because small sample sizes by discipline and grade level subgroups did not allow for reliably unbiased estimates of the treatment effect on students by grade level or discipline. Our findings supported the hypothesis that criteria-referenced formative assessment would increase students' achievement in the arts, as measured by performance-based assessments of knowledge and skill, even when controlling for pre-treatment measures and key demographic characteristics.

The small, positive effect of formative assessment found in this study is consistent with findings from other studies of *Arts Achieve* (e.g., Chen & Andrade, 2016; Mastrorilli, Harnett, & Zhu, 2014; Valle, 2015) and studies conducted in core content domains (Bennett, 2011). The results of this study suggest that student learning in the arts is measurably deepened when students know what counts, receive feedback from their teachers, themselves, and each other, and have opportunities to revise. The study supports Hattie and Timperley's (2007) claim that formative assessment contributes to learning because it helps students know where they are going, where they are now, and ways to close gaps in learning.

The results of this study also suggest that, through formative assessment, students can serve as useful resources for one another and can take ownership of their own learning (Valle, 2015; Wiliam & Thompson, 2007; see Figures 1 and 2). The students in this study engaged in frequent, formative peer and self-assessment according to checklists and rubrics, many of which they had co-created with their teachers. They did not grade themselves or each other—rather, they were taught to provide feedback according to the criteria for a task, with the explicit intention of helping themselves and their peers improve their work and deepen their learning. This study indicates that, under those conditions, students can effectively engage in formative assessment.

Despite the rigorous research design, data preparation, and analysis, this study has limitations, mainly due to constraints of the sample and insufficient evidence of reliability of some components of the arts achievement measures. Although the benchmark assessments were checked for content validity, and acceptable indices for internal consistency were obtained for most measures, the internal consistency of the visual arts measures for elementary and middle schools and for high school dance were low.

The inter-rater reliabilities for some of the tasks were also very low (k = .11 to k = 1.00). We included all of them in the analysis anyway for several reasons. First, of the 409 task criteria, only 55 (*i.e.,* 13.45%) had kappa values less than .40. Second, although a criterion of kappa of .40 or greater on a dichotomous variable is commonly considered acceptable in clinical settings (Sim & Wright, 2005), the choice of a minimum threshold value for acceptable inter-rater reliability is somewhat arbitrary—it depends on the context, the significance of the decisions to be made, and other factors. With the exception of music, measures of student learning in the arts are relatively less mature than those of core school subjects such as math and science. Although the BAA utilized many performance tasks that were authentic to the art forms being measured, the nature of creativity in the arts means it will take time for new measures in theater, dance, and visual arts education to meet high standards of psychometric quality.

In addition, as the number of scale categories increases, the possibility of disagreement between raters also increases. Given the four-level scale used in the Benchmark Arts Assessment scoring rubrics, as well as the challenge of assessing the arts, some low kappa values are to be expected. Considering that the Benchmark Arts Assessments are among the first measures of arts learning to be carefully developed by teams of experts through standardized procedures and the fact that there was evidence of the construct validity, or authenticity, of the tasks and rubrics, we included all the tasks in order to provide a comprehensive measure of arts learning.

Finally, the findings in this study are all conditional on the covariates that are included in our propensity score model. It is possible that any additional confounding variables might influence the results. Experiments that use random assignment at the student level are needed in order to further examine the effect of formative assessment on student performance in arts. Conducting formative assessment research in contexts dissimilar from New York City could also enhance the generalizability of the results.

Conclusion

This study is among the first to empirically investigate the effectiveness of criteria-referenced formative assessment in promoting students' learning in the arts, and one of the few large-scale experimental studies of formative assessment. Results from the pre-post randomized block design showed that criteria-referenced formative assessment had an overall small, positive effect on students' arts achievement. The findings provide evidence that by articulating success criteria and supporting students in providing feedback to themselves and each other, formative assessment can positively influence learning and achievement in the arts.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrade, H. G. (2001, April 18). The effects of instructional rubrics on learning to write. *Current Issues in Education* [On-line], *4*(4). Retrieved from http://cie.ed.asu.edu/volume4/number4

Andrade, H. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment.* New York: Routledge.

Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education, 17*(2), 199-214.

Andrade, H., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practices, 27*(2), 3-13.

Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice, 18*(1), 5-25.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5-31.

Brown, G. T., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 367-393). Los Angeles: SAGE.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*(3).

Chen, F. & Andrade, H. (2016). The impact of criteria-referenced formative assessment on fifth-grade students' theater arts achievement. *The Journal of Educational Research.* doi: 10.1080/00220671.2016.1255870

Colwell, R. (2004). Evaluation in the arts is sheer madness. *ARTSPRAXIS, 1*, 1 -12. Retrieved from http://steinhardt.nyu.edu/music/artspraxis

Diamond, A., & Sekhon, J. S. (2005). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*. Retrieved from http://escholarship.org/uc/item/8gx4v5qt#page-1

Englebright, K., & Mahoney, M. R. (2012). Assessment in elementary dance education. *Journal of Dance Education, 12*(3), 87-92.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234-249.

Harding, M. (2012). Assessment in the high school technique class: Creating thinking dancers. *Journal of Dance Education, 12*(3):93-98.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539-559.

Mastrorilli, T. M., Harnett, S., & Zhu, J. (2014). *Arts Achieve* impacting student success in the arts: Preliminary findings after one year of implementation. *Journal for Learning through the Arts, 10*(1).

Metis Associates (2015). *Arts Achieve* 2011-2014 internal consistencies. Unpublished report. New York : Author.

National Endowment for the Arts, (2012). *The arts and achievement in at-risk youth: Findings from four longitudinal studies (Research Report #55).* Washington, DC: J. S. Catterall, S. A., Dumais, & G Hampden-Thompson.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ross, J., & Starling, M. (2008). Self-assessment in a technology-supported environment: The case of grade 9 geography. *Assessment in Education: Principles, Policy and Practice, 15*(2), 183-199.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy, 85(3), 257–68. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15733050

Topping, K. (2013). Peer assessment as a form of formative assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 395-412). Los Angeles: SAGE.

Valle, C. (2015). Effects of criteria-referenced formative assessment on achievement in music (Doctoral dissertation). Retrieved from ProQuest. (3740126)

Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment.* New York: Routledge.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), The Future of Assessment: Shaping Teaching and Learning (pp. 53–82). Mahwah, NJ: Erlbaum.

Table 1

Student Demographic Information for the Complete Sample after Imputation

	Overall (n=2219)		**Control (n=1608)**		**Treatment (n=611)**	
	N	**% of Total**	**N**	**% of Total**	**N**	**% of Total**
Discipline						
Dance	468	21.09	400	18.03	68	3.06
Music	608	27.40	541	24.38	67	3.02
Theater	530	23.88	306	13.79	224	10.09
Visual Arts	613	27.63	361	16.27	252	11.36
English Language Learner (ELL)						
Not ELL	1945	87.65	1402	63.18	543	24.47
ELL	274	12.35	206	9.28	68	3.06
Ethnicity						
American Indian or Alaskan Native	13	0.59	3	0.14	10	0.45
Asian or Pacific Islander	444	20.01	317	14.29	127	5.72
Black, not of Hispanic Origin	580	26.14	457	20.59	123	5.54
Hispanic	742	33.44	563	25.37	179	8.07
White, not of Hispanic Origin	433	19.51	262	11.81	171	7.71
Multiracial	6	0.27	5	0.23	1	0.05
Parents declined to declare	1	0.05	1	0.05	0	0.00
Free or Reduced Lunch (FRL)						
Not FRL	972	43.80	710	32.00	262	11.81
FRL	1247	56.20	898	40.47	349	15.73
Gender						
Male	949	42.77	670	30.19	279	12.57
Female	1270	57.23	938	42.27	332	14.96
School Level						
Elementary	1072	48.31	696	31.37	376	16.94
Middle	730	32.90	553	24.92	177	7.98
High	417	18.79	359	16.18	58	2.61
Special Education						
Not Special Ed	2076	93.56	1529	68.90	547	24.65
Special Ed	143	6.44	79	3.56	64	2.88

Table 2

Student Performance Measures, Complete Sample after Imputation

		Overall (n=2219)	**Control (n=1608)**	**Treatment (n=611)**
Average Daily Attendance	M	94.97	95.02	94.82
	SD	4.92	4.95	4.86
ELA Achievement	M	667.54	667.35	668.04
	SD	25.51	25.09	26.62
Writing	M	59.16	58.56	60.74
	SD	20.42	20.51	20.12
Math Achievement	M	696.88	698.15	693.53
	SD	34.08	34.54	32.62
2011 Pre-assessment	M	55.64	55.84	55.1
	SD	17.31	17.44	16.98
2012 Post-assessment	M	61.66	61.19	62.89
	SD	18.01	18.42	16.81

Table 3

Internal Consistencies for Fall 2011 and Spring 2012 Benchmark Arts Assessments

	Fall 2011		Spring 2012	
	# of Items	α	# of Items	α
Dance				
Elementary	16	.86	16	.86
Middle	15	.85	15	.88
High	19	.47	19	.84
Music				
Elementary: Voice	15	.77	15	.72
Elementary: Instrumental	15	.76	15	.76
Middle	23	.88	23	.86
High	26	.93	26	.83
Theater				
Elementary: Playwriting	15	.81	12	.85
Elementary: Costume Design	16	.73	13	.83
Middle: Acting - Directing /Actors	9	.88	9	.81
Middle: Acting – Design	9	.84	7	-
Middle: Musical Theater - Directing/Actors	9	.81	7	.77
Middle: Musical Theater - Design	9	.74	7	-
Middle: Playwriting & Directing/ Actors	9	.90	7	.78
Middle: Playwriting & Design	9	.83	7	-
High: Acting: Character	9	.90	-	-
High: Acting: Design	9	.88	11	.77
Visual Arts				
Elementary	32	.18	29	.85
Middle	28	.46	25	.88
High	18	.87	18	.87

Note: There are no alphas for Spring 2012 Acting-Design, Musical Theater-Design, and Playwriting & Design because only 0-1 students completed the items.

Table 4

Number (n) of student pairs by discipline and educational level after matching

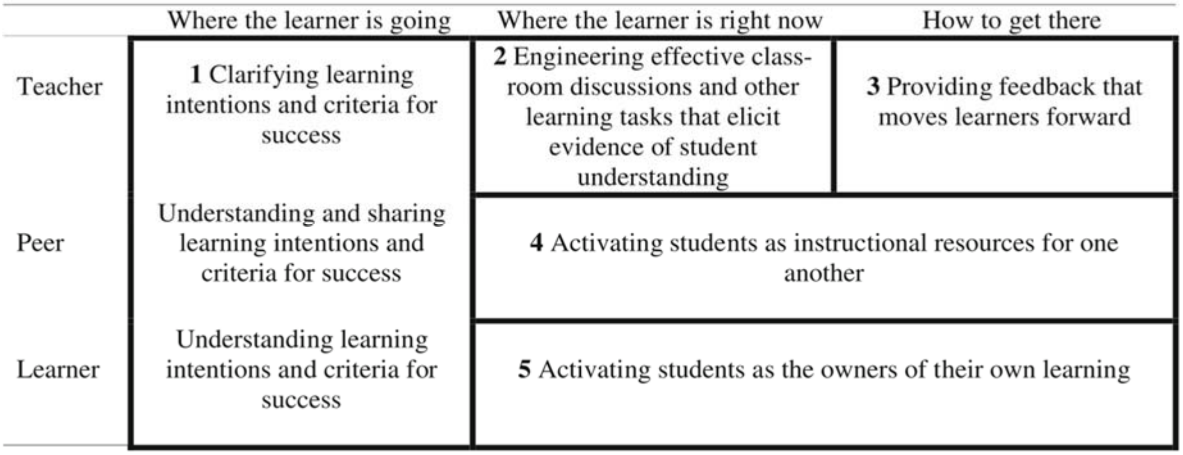
	Dance	Music	Theater	Visual Arts	Total
Elementary	25	67	144	140	376
Middle	43	0	58	76	177
High	0	0	22	36	58
Total	68	67	224	252	611

Note: ns represent the numbers of matched pairs of students.

Table 5

Student Demographic Information for the Matched Sample

	Overall **(n=1222)**		**Control** **(n=611)**		**Treatment** **(n=611)**	
	N	**% of Total**	**N**	**% of Total**	**N**	**% of Total**
English Language Learner (ELL)						
Not ELL	1083	88.63	540	44.19	543	44.44
ELL	139	11.37	71	5.81	68	5.56
Ethnicity		0.00				
American Indian or Alaskan Native	10	0.82	0	0.00	10	0.82
Asian or Pacific Islander	221	18.09	94	7.69	127	10.39
Black, not of Hispanic Origin	307	25.12	184	15.06	123	10.07
Hispanic	344	28.15	165	13.50	179	14.65
White, not of Hispanic Origin	335	27.41	164	13.42	171	13.99
Multiracial	3	0.25	2	0.16	1	0.08
Parents declined to declare	2	0.16	2	0.16	0	0.00
Free or Reduced Lunch (FRL)		0.00				
Not FRL	521	42.64	259	21.19	262	21.44
FRL	701	57.36	352	28.81	349	28.56
Gender		0.00		0.00		0.00
Male	584	47.79	305	24.96	279	22.83
Female	638	52.21	306	25.04	332	27.17
Special Education		0.00				
Not Special Ed	1092	89.36	545	44.60	547	44.76
Special Ed	130	10.64	66	5.40	64	5.24



	Where the learner is going	Where the learner is right now	How to get there
Teacher	**1** Clarifying learning intentions and criteria for success	**2** Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding	**3** Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	**4** Activating students as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	**5** Activating students as the owners of their own learning	

Figure 1. Formative Assessment Framework (Wiliam & Thompson, 2007).

	Where the learner is going	Where the learner is right now	How to get there
Teacher	Clarifying learning intentions and criteria for success	Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding	Providing feedback that moves learning forward
Peer	Understanding and sharing learning intentions and criteria for success	Activating students as instructional resources for one another	
		Assessing the quality of another's work in relation to criteria for success	Providing feedback to a peer about how to move learning forward
Learner	Understanding learning intentions and criteria for success	Activating students as the owners of their own learning	
		Assessing the quality of own work in relation to criteria for success	Providing feedback to self about how to move learning forward

Figure 2. Modified Formative Assessment Framework (Valle, 2015).

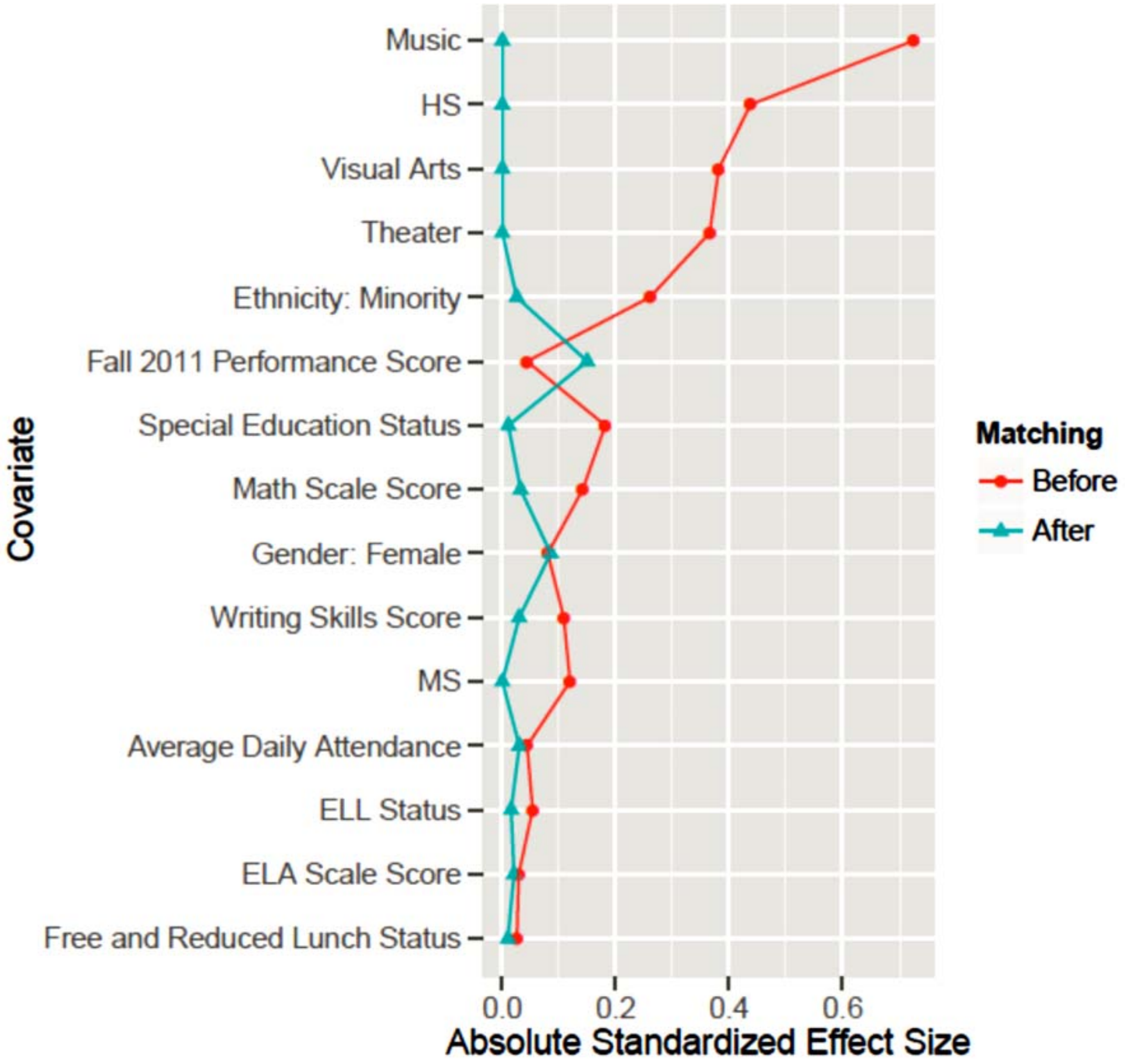
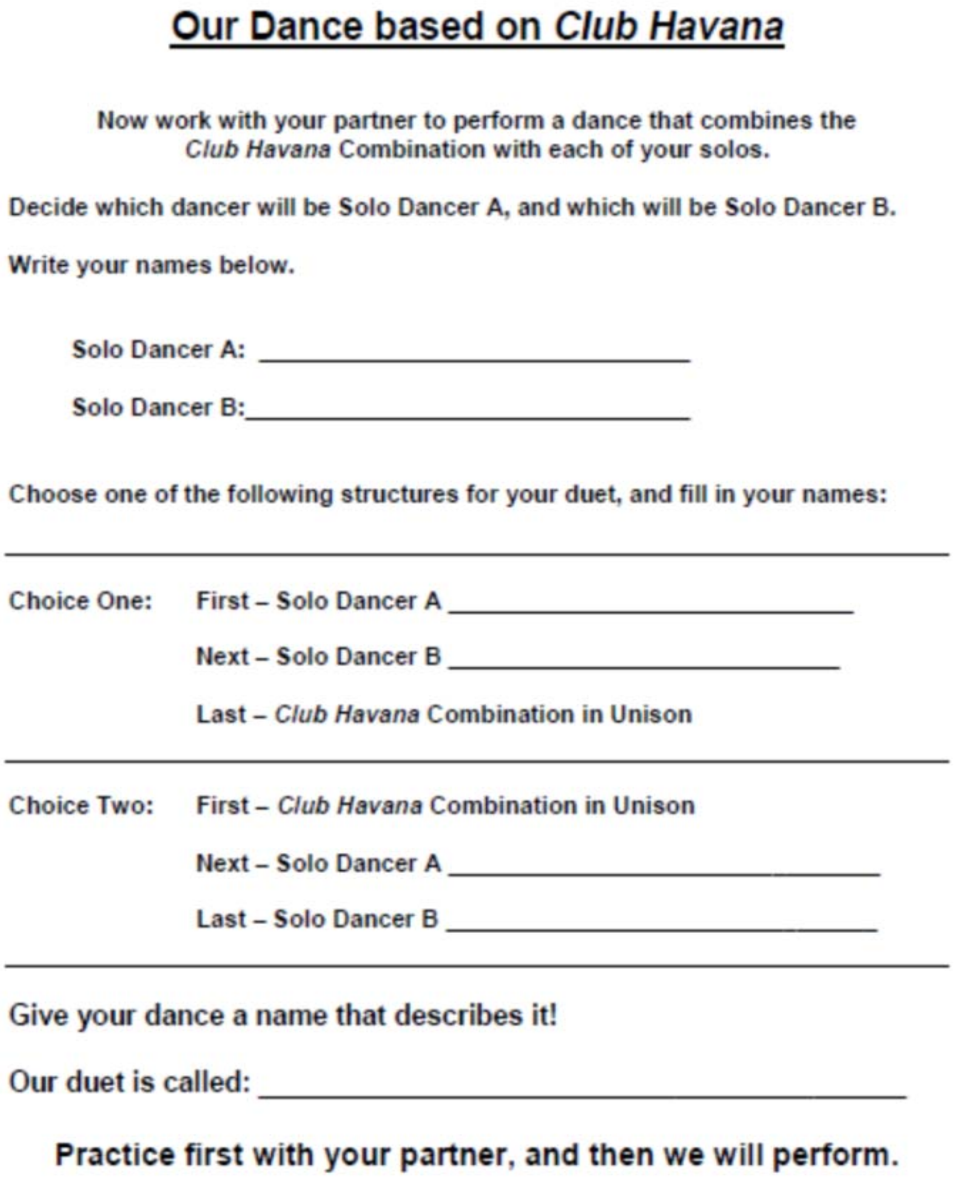


Figure 3: Balance of covariates in the propensity score model before and after matching.

Appendix A

Sample Grade 5 Dance Performance Task



Our Dance based on *Club Havana*

Now work with your partner to perform a dance that combines the *Club Havana* Combination with each of your solos.

Decide which dancer will be Solo Dancer A, and which will be Solo Dancer B.

Write your names below.

Solo Dancer A: ______________________________

Solo Dancer B:______________________________

Choose one of the following structures for your duet, and fill in your names:

Choice One:	**First – Solo Dancer A** ____________________________
	Next – Solo Dancer B ___________________________
	Last – *Club Havana* Combination in Unison
Choice Two:	**First – *Club Havana* Combination in Unison**
	Next – Solo Dancer A ______________________________
	Last – Solo Dancer B ______________________________

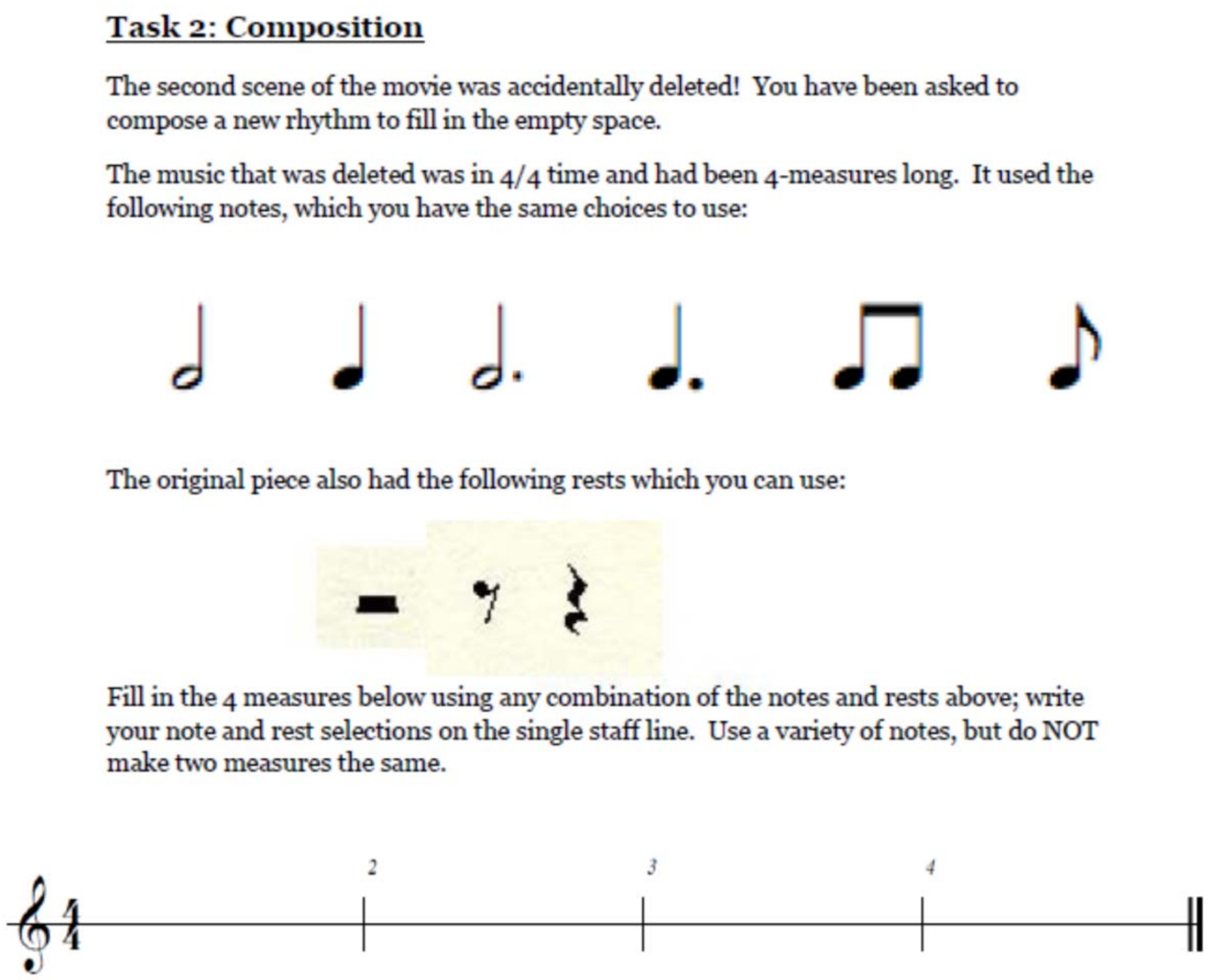
Give your dance a name that describes it!

Our duet is called: ____

Practice first with your partner, and then we will perform.

Appendix B

Sample Grade 8 Music Performance Task



Task 2: Composition

The second scene of the movie was accidentally deleted! You have been asked to compose a new rhythm to fill in the empty space.

The music that was deleted was in 4/4 time and had been 4-measures long. It used the following notes, which you have the same choices to use:

𝅗𝅥 ♩ 𝅗𝅥. ♩. ♫ ♪

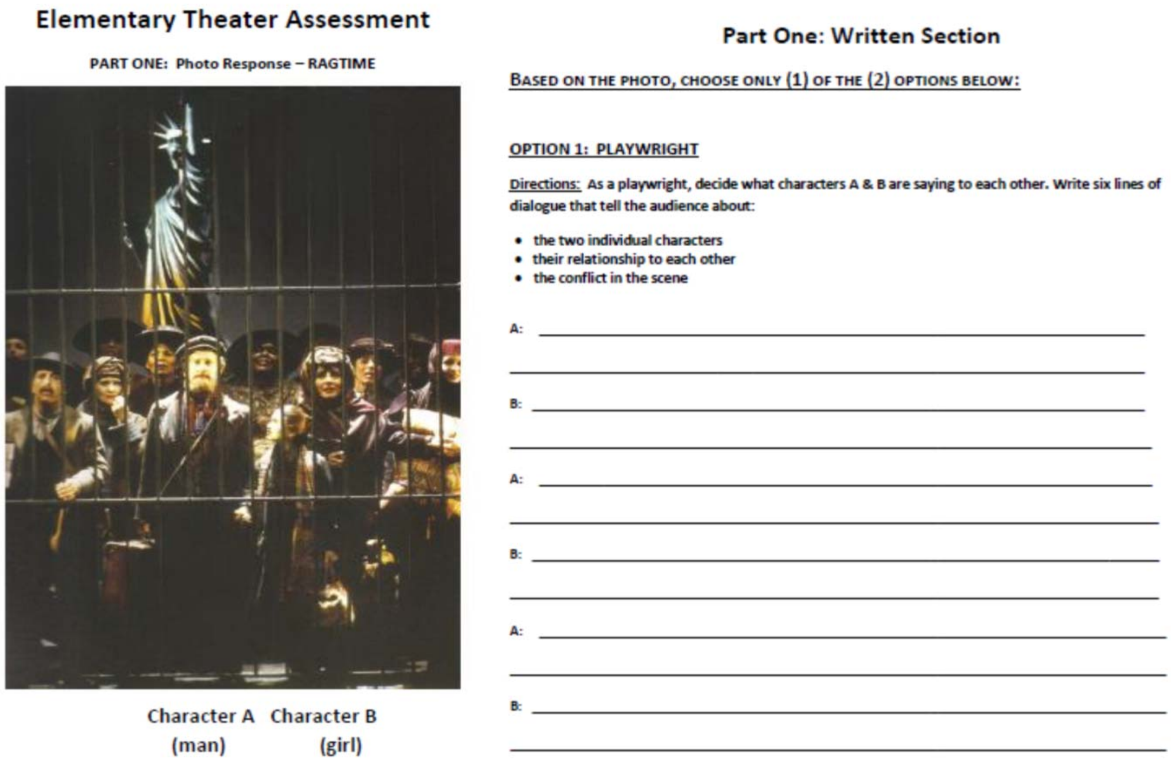
The original piece also had the following rests which you can use:

𝄼 𝄾 𝄽

Fill in the 4 measures below using any combination of the notes and rests above; write your note and rest selections on the single staff line. Use a variety of notes, but do NOT make two measures the same.

Appendix C

Sample Grade 5 Theater Performance Task



Elementary Theater Assessment

PART ONE: Photo Response – RAGTIME

Character A (man) Character B (girl)

Part One: Written Section

BASED ON THE PHOTO, CHOOSE ONLY (1) OF THE (2) OPTIONS BELOW:

OPTION 1: PLAYWRIGHT

Directions: As a playwright, decide what characters A & B are saying to each other. Write six lines of dialogue that tell the audience about:

- the two individual characters
- their relationship to each other
- the conflict in the scene

A: ______________________________

____

B: ______________________________

____

A: ______________________________

____

B: ______________________________

____

A: ______________________________

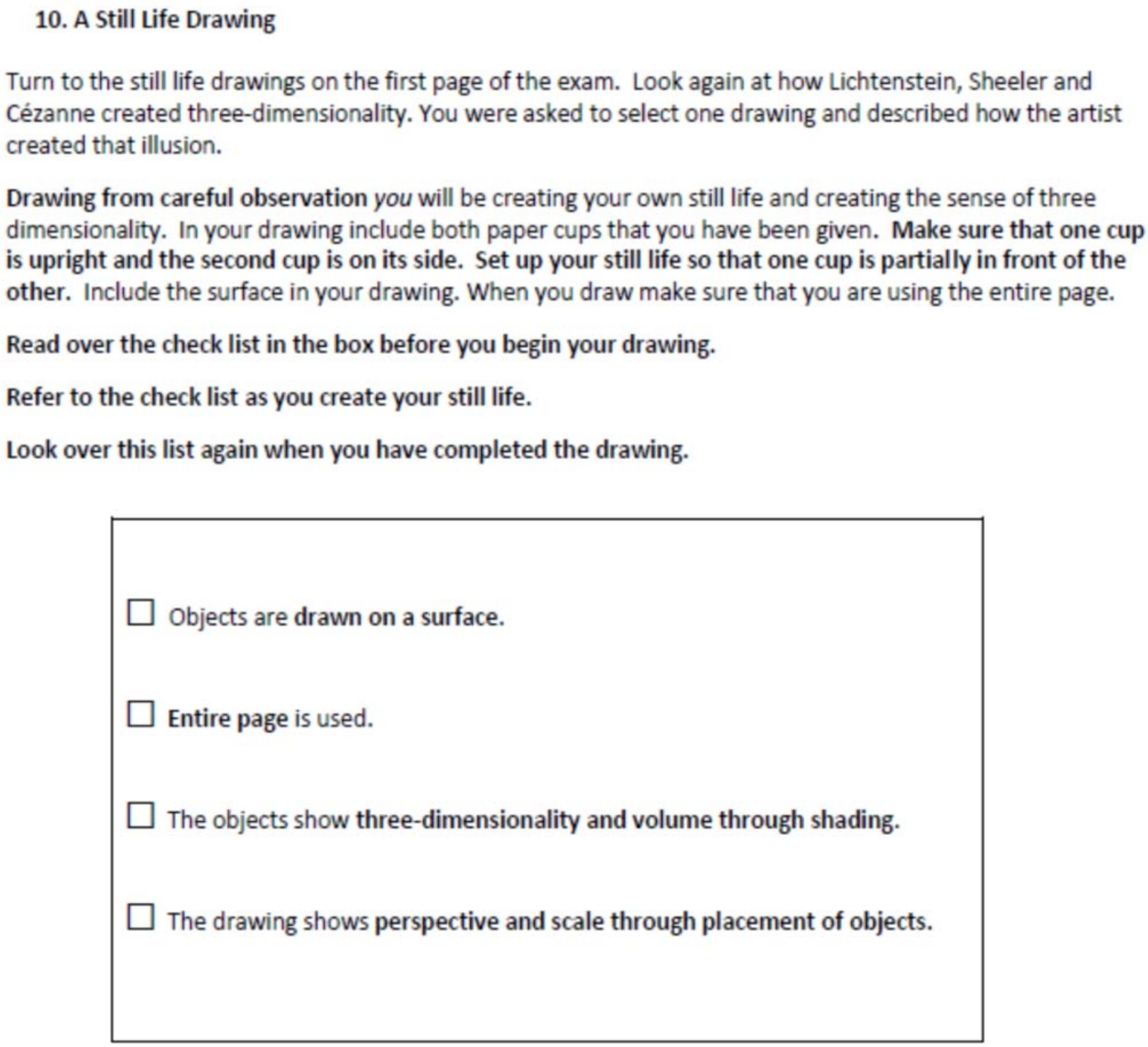
____

B: ______________________________

____

Appendix D

Sample Grade 12 Visual Arts Performance Task



10. A Still Life Drawing

Turn to the still life drawings on the first page of the exam. Look again at how Lichtenstein, Sheeler and Cézanne created three-dimensionality. You were asked to select one drawing and described how the artist created that illusion.

Drawing from careful observation *you* will be creating your own still life and creating the sense of three dimensionality. In your drawing include both paper cups that you have been given. **Make sure that one cup is upright and the second cup is on its side. Set up your still life so that one cup is partially in front of the other.** Include the surface in your drawing. When you draw make sure that you are using the entire page.

Read over the check list in the box before you begin your drawing.

Refer to the check list as you create your still life.

Look over this list again when you have completed the drawing.

□ Objects are **drawn on a surface.**

□ **Entire page** is used.

□ The objects show **three-dimensionality and volume through shading.**

□ The drawing shows **perspective and scale through placement of objects.**