



Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Lights up! Assessing standards-based performance skills in drama education

Kylie Pepler^{a,1,2}, Sophia Bender^{b,3}, Anthony Phonethibsavads^{a,4}, Nickolina Yankova^{a,*},
Molly Stewart^b

^a School of Education, University of California, Irvine, Irvine, CA, USA

^b School of Education, Indiana University Bloomington, Bloomington, IN, USA

ARTICLE INFO

Keywords:

Drama assessment
Drama/ theatre standards
Scoring rubric
Rubric development

ABSTRACT

With the recent release of the National Core Arts Standards (NCAS), educators need new consistent, fair assessments of drama learning. As an initial starting point, this paper reports on the creation of the LATA Drama Performance Rubric, a standards-based assessment to measure learning occurring under real drama classroom conditions that we hope evaluators will find to be useful. A widespread group of drama instructors coordinated with researchers to create a rubric containing four categories: (1) Diction and Volume; (2) Movement and Gesture; (3) Group Coordination; and (4) Stage Presence. Field testing of the instrument with 97 students in the treatment group and 80 students in the control group demonstrated its ability to distinguish between fourth-grade classrooms that had and had not received long-term drama instruction. Reliability, validity, and NCAS alignment are discussed, along with limitations and future recommendations.

1. Introduction

In recent decades, evaluation has taken a central role in educational settings. It has primarily served two functions, accountability and amelioration, with the former using assessment of student academic performance to evaluate the effectiveness of educational programs, and the latter seeking improvement of existing programs (e.g., Love, 2010; Mathison, 2010). Such valuational efforts have aided funders and the federal government take stock of ongoing efforts and consider avenues for improvement that would be of benefit to students in providing more efficacious and higher quality education. Though there are divergent perspectives on what is to be considered an outcome regarding student assessments of learning, what has largely been prioritized are cognitive and socioemotional measures, to the detriment of other assessment efforts despite current research questioning knowledge as a primary outcome (Schwartz & Arena, 2013). In the arts, and drama/theatre especially, there has been a need for reliable and valid assessments for learning that evaluate arts-related outcomes (Haanstra et al., 2015), especially at the elementary level (e.g., Omasta et al., 2021).

Toward ensuring a quality arts education across grade levels, the National Coalition for Core Arts Standards (NCCAS) revisited the National Core Arts Standards (NCAS) in 2014, with the aim of providing a guiding framework for arts educators, evaluators and other stakeholders to inform teaching and assessment practices (National Coalition for Core Arts Standards, 2014). The NCAS is rooted in four domains – creating, performing, responding, and connecting – and anchor standards by grade level from pre-k to high school across arts disciplines (visual arts, music, dance, theatre). Since the introduction of the NCAS, twenty-seven states have passed new or revised standards and twenty-two states have added standards for the emergent discipline of media arts (NCCAS, 2019).

Historically, there has not been one “correct” pedagogical model for drama education because much of it is dependent on the social, political, and cultural climate of the times (Bolton, 2007). Consequently, there has also been variation in the types and rigor of drama assessment tools that have been developed and utilized, especially at the elementary level. Coupled with the drama field’s nascent history in traditional classroom contexts, it is not surprising that published research-based measures for

* Correspondence to: 7125 Palo Verde Rd, Irvine, CA 92617, USA,
E-mail address: nyankova@uci.edu (N. Yankova).

¹ <https://orcid.org/0000-0002-5472-4974>

² Twitter: @drpepler

³ <https://orcid.org/0000-0001-8039-4414>

⁴ <https://orcid.org/0000-0001-5206-8195>

⁵ <https://orcid.org/0000-0002-7552-527X>

<https://doi.org/10.1016/j.stueduc.2023.101259>

Received 31 July 2021; Received in revised form 9 June 2022; Accepted 17 March 2023

Available online 29 March 2023

0191-491X/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

drama learning that have been shown to be reliable and valid are few and far between. One such example is the Preschool Theatre Arts Rubric (PTAR) that Susman-Stillman and colleagues (2018) created in accordance with the NCAS and state standards to identify important competencies for preschool-age children that could be developed through theatre arts skills. Furthermore, the NCCAS has published a number of model cornerstone assessments aligned with standards that provide guidance on administering the assessment (e.g., strategies for embedding in instruction, task specific rubrics, differentiation strategies, etc.), and a comprehensive list of learning goals including items related to new knowledge, skills, and vocabulary (NCCAS, 2016). However, these assessments tend to evaluate a final product in drama/theatre and not so much drama learning at a given point in time. Towards filling this research gap, what is needed is a measure of the extent to which kids are learning drama where learning is conceptualized as change over time at the individual level. Such a measure would be especially useful for external assessors in determining a drama program's quality in teaching drama as opposed to evaluating its efficacy based on non-arts related learning outcomes.

As an initial starting point, this paper reports on the construction and design of the Learning and Achieving through the Arts (LATA) Drama Performance Rubric, an assessment rubric for select drama skills that we hope evaluators will find to be useful. The rubric finds alignment with the NCAS at the early elementary level and includes the following categories: Diction and Volume (Diction/Volume), Movement and Gesture (Movement/Gesture), Group Coordination, and Stage Presence. The goal of the rubric is to reliably and fairly show learning of standards-based drama skills over time, without assessing talent or performance ability (Oreck et al., 2003). Standards tend to value what happens routinely and under natural settings. Thus, the rubric builds on what is already happening in the classroom by evaluating a child's performance during a common warm-up activity. As such, the measure aims to be an unobtrusive if not invisible part of a classroom's routine in an evaluation process. The measure aims to assess to what extent kids are learning drama toward understanding the efficacy of drama programs, but it is by no means exhaustive in terms of measuring drama skills. Given that there are not many measures that examine change over time in drama learning, the present one is promising as it does not "black box" drama learning. We conducted a quasi-experimental pre-post study with six 4th grade classrooms (97 students in the treatment group and 80 students in the control group) to provide initial evidence of the rubric's validity and reliability. We end with a discussion of limitations of the instrument and recommendations for future work.

2. Drama assessments: an overview of the literature

2.1. Rationales for assessing learning and skills in drama

Much of the prior work on drama-related assessment has focused on the ways that learning in the arts may transfer to non-arts learning (Bransford & Schwartz, 1999; Deasy, 2002; Hardiman et al., 2014; Ludwig et al., 2017; Winner & Hetland, 2004), child development (Foster and Marcus Jenkins (2017); Mages, 2018), and socioemotional skills (Gallagher & Service, 2010; Li et al., 2015). Researchers have studied classroom drama and its association with performance in other academic competencies and fields of study such as reading comprehension (e.g., Fleming et al., 2004; Kelner & Flynn, 2006; Rose et al., 2000), English Language Arts (e.g., Peppler, Catterall, & Bender, 2015; Greenfader et al., 2015; Ragpot, 2011; Walker, McFadden, et al., 2011; Walker, Tabone, et al., 2011), math (e.g., Fleming et al., 2004; Walker, Tabone, et al., 2011), social studies (e.g., Walker, McFadden, et al., 2011), and history (e.g., Kisida et al., 2020). The effects of drama programs seem to be generally positive on these other subject areas, but most studies of this type assess only the non-arts learning, and do not measure improvement in drama skills. Most recently, studies targeting pre-k and elementary contexts have focused on drama's positive effects

on reducing instances of aggression (Korošec & Zorec, 2020); improving self-concept (DeBettignies & Goldstein, 2020); supporting the development of 21st century skills such as language, collaboration, and creative problem solving (e.g., Brown, 2017); and easing the transition from primary to secondary schooling (Barlow, 2020).

In contrast, the *National Coalition for Core Arts Standards (2020)* highlights the benefits of the arts in their own right, independent of implications for non-arts learning; in that sense, art is perceived as a valuable form of communication and creative personal realization that is able to bridge one's culture and history and contribute to one's well-being, further facilitating community engagement. Therefore, it is important to include art in the curriculum for its own sake, which necessitates the construction of its own assessment measures that evaluate change over time in the context of drama learning exclusively and specifically. For a more comprehensive review of these issues relating to the multiple purposes of drama education, see *Weltsek et al. (2014)*.

2.2. Use of standards-based assessment rubrics in drama education research

In drama education research, there has been a noticeable lack of scholarly work around research-based assessment instruments (e.g., Haastra et al., 2015; Omasta, 2021). In a comprehensive review of the literature, Haastra and colleagues (2015) found that only 3% of the 153 arts-assessment articles published in peer-reviewed journals since 2000 and examining assessment instruments have featured drama assessment tools. Though subjectivity is inevitably present in evaluating creative processes in drama/theatre, assessment is still an important part of drama education and is often a contested space where transparency is needed to make drama assessment more accessible to experts and non-experts alike (Jacobs, 2022). Academic standards require well-aligned assessments for evaluators and educators to determine whether learning has occurred according to the criteria in the standards. With drama education's innate subjectivity and varying quality of instruction by teacher (Oreck et al., 2003), rubrics have the potential to ensure a higher minimum quality across settings.

In the arts, assessment instruments in the form of observation rubrics are particularly useful, where multiple media and performances may be incorporated in place of, or in addition to, written assessments. Rubrics aligned with standards provide teachers with ways to assess students reliably, track growth over time, and compare students or classes on various components of the standards, in order to target further instruction. Rubrics are a quick way to lay out clear expectations for students to follow and to evaluate learning, and they can be readily adapted to suit many assessment purposes, including 'diagnostic, formative, summative, authentic, or traditional' (Van de Water et al., 2015). The one rubric that has been explicitly validated for the NCAS and published in a peer-reviewed journal is the one by Susman-Stillman and colleagues (2018), which used NCAS as well as state standards to identify important theatre competencies for preschool-age children in the context of storytelling and storyacting, zooming in on the following: independence in role play, use of face and gesture, focus/persistence, collaboration, and theatricality. The authors tested the instrument for validity and reliability, finding it was suitable to be used with preschool-age children (Susman-Stillman et al., 2018).

Most published drama assessment instruments tend not to discuss alignment with standards. For instance, Oreck and colleagues (2003) developed an assessment instrument that objectively, validly, and reliably assessed talent in the performing arts—i.e., whether students are likely to succeed in advanced art instruction, not necessarily how well they perform in the moment. This rubric was tested over several years at three schools in New York City and two schools in Ohio. It involves a checklist of performance-related behaviors that observers can mark as present or absent in learners at the time the rubric is administered. However, the checklist's criteria are not based on existing standards, and it is meant to be a one-time evaluation of a young person's

competence in the performing arts skills rather than a measure to show learning over time. Kelner and Flynn (2006) provide samples of drama assessments in the form of observation protocols, reflective discussion questions for formative assessment with students, written assessments, and a rubric-like checklist. While the authors list many performance criteria, they are again not explicitly linked to specific standards, nor do they provide evidence that the sample assessments are valid, equitable, or reliable. Korkut (2018) developed and piloted an assessment instrument related to drama, but it was a rubric to assess the creativity of pre-service teachers' drama lesson plans, not an assessment of students' drama skills, that was also not aligned with specific standards.

As for arts assessments that have been aligned to standards, Chen et al. (2017) studied criteria-referenced formative assessment in the arts (including theatre), which were aligned with New York City's guidance on arts education and the Common Core State Standards for language arts. Data were collected in 2011 and 2012, predating publication of the NCAS standards. Another study (Lin, 2013) reported on the creation of a rubric for assessing drama performances based on Taiwan's national Arts and Humanities standards. By aligning to standards, these instruments gained utility for use anywhere the standards are in place. Especially at the elementary level, there is a need for additional research into drama assessment instruments and practices toward providing better support for teachers and a higher quality drama education for students. Most recently, Omasta and colleagues (2021) undertook a phenomenological study with elementary school drama teachers to understand their experience with assessment, with policy implications at the school, district, and state levels. Findings discuss some of the challenges at present of drama assessment practices in elementary classrooms (e.g., constraint of time in the regular assessment of authentic drama learning; achieving balance between granting teacher agency and providing support in navigating expectations by various stakeholders). Though regular assessment can be burdensome, it is imperative to align learning goals with standards and administer assessments as evidence of learning to secure funding and ensure better quality arts education in elementary drama classrooms in particular.

In the current standards-based environment in the U.S., assessments not based on standards are difficult for teachers to adopt because standards are designed to guide teachers' instruction and assessment. In addition, assessments that are not aligned with standards may lead to issues with reliability and validity (Baptiste, 2008; Jacobs, 2016; Oreck et al., 2003). Non-aligned arts assessments are commonly based on subjective judgments of "artistic skill" and "aesthetics" (Baptiste, 2008). While these subjective forms of assessment can help form a complete picture of a student's drama skills, they should not necessarily be used alone (nor should standards-based rubrics be used alone). However, with published standards-based drama assessment instruments that have been tested for reliability and validity few and far between, many teachers must design their own assessments. Many resources exist that provide teachers with guidelines on how to create their own drama assessments (e.g., Kelner & Flynn, 2006; Van de Water et al., 2015), but these place heavy burden on teachers and do not address the issues of consistency, validity, reliability, generalizability, fairness, and alignment with standards. See Kelner and Flynn (2006) and Meyer (2016) for discussions of additional assessment methods in drama.

2.3. Construction and design of the LATA Drama Performance Rubric

The present study attempts to begin filling this gap by introducing the LATA Drama Performance Rubric, an assessment instrument for a select few drama skills, aimed at the early elementary level and found in its entirety in Appendix A. It is an initial exploration of a tool we hope evaluators and other interested stakeholders will find to be useful. The proposed rubric is based on a common drama warm-up exercise (i.e., the name-and-movement game). As such, it lends itself to a natural pre and post measure in drama toward drawing comparisons between the start and end of an intervention. Through clearly defined descriptions for

each level of performance, the rubric holds the potential to make judgments of student performance more "objective" and evaluate student learning at a point in time. It further aims to evaluate drama learning over time, rather than a final product, which few rubrics at the elementary level have done (Omasta et al., 2021). The rubric uses a 5-point scale, intentionally created to be ordinal rather than interval. To provide evidence of the instrument's validity and reliability, we conducted a quasi-experimental pre-post study, which we discuss in the following two sections.

The rubric used for data collection and analysis in this paper was developed over several years with teaching artists in the Los Angeles area, with the latest version completed in 2011. Data were collected in 2012 and 2013, and we analyzed these results for pre-post growth. The rubric scoring categories were collaboratively constructed with four drama teaching artists across four different art programs for youth. All teaching artists were chosen as rubric design consultants based on their joint expertise in both the art and teaching of drama, helping to ensure that the categories they identified would have content validity. In constructing the rubric, we aimed to find alignment between what is valued by standards, what skills are taught in the classroom and what drama skills the rubric should focus on. Earlier versions of the rubric were created and piloted in 2006 (Pepler & Catterall, 2006) and 2009 (Pepler, Catterall, & Feilen, 2009).

The four categories identified by the teaching artists were diction and volume, variety of improvised movement, group coordination, and use of neutral position. While these competencies were pertinent to the name-and-movement game, they also represented drama skills and learning outcomes teaching artists hoped students would take away upon completion of their drama classes. For the name-and-movement game, students learn to speak their name loudly and clearly (diction and volume), but this skill is equally applicable to onstage performances. Additionally, selecting a movement that is visible, reproducible, and expressive is an important skill that is learned in the activity that is also transferable to other contexts (e.g., portraying a character on stage through expressive movement). By carefully imitating all of their peers' names and movements, students exhibit cooperation and group coordination, showing respect for both the group and activity, which is a universal skill across contexts. Lastly, maintaining a neutral position – outside of staying focused and alert – teaches students to be a respectful, attentive audience for their peers.

The following section depicts the rubric's alignment to the NCAS with a brief description of each category and its importance. Category names reflect the original names as identified by the four drama teaching artists in collaboration with the research team. To find better alignment with the NCAS, we propose revised category names for the next iteration of the rubric. This can also be found in Appendix B. Under the NCAS 2014, the rubric is most suitable for use in grades K-3 and targets primarily two out of the four domains put forth by the NCAS: creating and performing.

2.3.1. Diction and volume (Diction/Volume)

This category refers to how loudly and clearly a performer speaks, an essential skill in drama performance, especially at the elementary level (Linklater, 2006). It corresponds to the NCAS's "Creating" standard 3.1.2.b for second grade (TH:Cr3.1.2.b): "Use and adapt sounds and movements in a guided drama experience" and "Performing" standard 4.1.3.b for third grade (TH:Pr4.1.3.b): "Investigate how movement and voice are incorporated into drama/theatre work." For better alignment with the NCAS, we propose this category be renamed to Expression through sound and voice.

2.3.2. Movement and gesture (Movement/Gesture)

This category entails the expressiveness and originality of participants' movements. Imagination is an essential skill in drama and this category reflects such an attention to originality and imaginativeness (Johnstone & Wardle, 2012). It aligns with the NCAS's "Creating"

standard 2.K.b for kindergarten (TH:Cr2.K.b): “With prompting and support, express original ideas in dramatic play or a guided drama experience”; this category also captures aspects of standards TH:Cr3.1.2.b and TH:Pr4.1.3.b covered above under “Diction and volume”. For better alignment with the NCAS, we propose this category be renamed to Expression through movement.

2.3.3. Group coordination

This category involves the participants’ cooperation within the activity, including following directions and responding appropriately to their classmates’ actions, an important skill for engaging in drama activities (Hagen & Frankel, 1973; Somers, 2005). Originally called Teamwork, we renamed the category to Group coordination so that it better captures the essence of what is evaluated. This category aligns with the NCAS’ “Creating” standard 2.7.b for seventh grade (TH:Cr2.7.b): “Demonstrate mutual respect for self and others and their roles in preparing or devising drama/theatre work” and the more general “Performing” standard 5.1.3 for third grade, TH:Pr5.1.3.a: “Participate in a variety of physical, vocal, and cognitive exercises that can be used in a group setting for drama/theatre work.”

2.3.4. Stage presence/Neutral position



The professional teaching artists whom we consulted for the design of the rubric felt strongly that maintaining neutral position (i.e.,

standing facing forward, arms at sides) was an important skill for students in drama activities to demonstrate. This posture suggests that an actor is not preoccupied with other thoughts and is focused, which allows the actor to react at any moment (Hagen & Frankel, 1973; Somers, 2005). Much like many drama warm-ups, this particular activity requires students to acknowledge and respond to their classmates performing the exercise with them, so it is necessary to stay alert and attentive in order to participate in the activity. Additionally, the hands positioned at sides free the actor to quickly perform any gesture, in contrast to having arms crossed, hands in pockets, etc., which prevents the actor from being prepared to act quickly. This category does not presently align with the NCAS standards, so we include it as a behavioral expectation. Future revisions of the rubric might consider a different skill that is in better alignment with current standards.

2.4. Sample student scoring

In Table 1 below we present the case of a student we call “Lucas” (a pseudonym) who was evaluated using the LATA Drama Performance Rubric at two time points: at the beginning and end of the LATA drama intervention. To illustrate the scoring process, we present a breakdown of his scores for each category with a brief justification for each score. Lucas was one of the students who showed the greatest improvement and achieved one of the highest scores in the post assessment, with his

Table 1
Pre and Post Scores of an Example Student’s Name-and-Movement Performance, Using the LATA Drama Performance Rubric.

	Pre Score	Post Score
Diction and Volume Movement and Gesture	Lucas spoke his name clearly, but not loudly enough for the stage, so he received a score of 3 for Diction and Volume. 	Lucas spoke his name clearly and loudly enough that he would definitely be able to be heard on a stage, so he received a score of 5. 
Group Coordination	Lucas did not imitate his neighbors in the pre, except for, at one point, moving his arms slightly, but this slight movement did not match his neighbor’s movement at all. He thus received a 1 for imitation.	Lucas’s best imitation was of the neighbor to his left whose turn came after his. Lucas imitated this neighbor’s dancing shuffle with attention to specific details in the movement. However, Lucas did not show as much enthusiasm as his neighbor did in his movement, so Lucas lost a point and received a 4.
Stage Presence/Neutral Position	While Lucas did remain in the neutral position—facing forward, standing straight, hands at side—he neglected to imitate his neighbors, indicating a lack of presence in the moment. He was ready to perform his own movement when his turn arrived, however, so he received a score of 2 for Stage Presence.	Lucas paid attention to the activity, and was ready to act when it was time to perform his own movement and time to imitate his neighbors. He remained in the neutral position most of the time, only deviating slightly when he briefly held his hands together, and later adjusted his clothing. Thus he received a score of 4 for Stage Presence.
Total Score	Lucas scored 9 points on Pre-Test	Lucas scored 18 points on Post-Test

overall score improving from 9 in the pre to 18 in the post. Although we were only able to use still images to convey his performance, the change in dynamism is still visible.

2.5. Adaptability of the Rubric

The purpose and expectation of a given drama activity will define the specific scoring criteria for each category. Although it is possible to assess performance ability without specific criteria (Pepler, Catterall, & Bender, 2015), the inclusion of clearly defined categories allows for evaluators, researchers, and educators to identify specific behaviors that may be more closely connected to differences in learning and creativity. This aspect of the rubric allows it to be flexible and adaptable to the specific teachers' and classrooms' needs. For instance, if the activity is the performance of a new monologue, some adjustments would need to be made to the rubric's criteria to incorporate aspects that account for memorization and performance quality. As opposed to simply evaluating if a performer speaks loudly and clearly in the Diction/Volume category, a consideration of the performer's inflection and memorization quality might be incorporated. One drawback of this is that the rubric should not be used to compare students' performance of different activities, but rather the same activity at different points in time. The rubric is also not intended to compare students to each other, as the ordinal rating structure allows for a variety of expressions that may be quantified by the same ratings (for example, vocal expression which is loud but not clear, or clear but not loud, would be rated the same in this rubric). Additionally, because drama performances are inherently sensitive to changes in contextual factors, such as environments or individual anxieties, we advise observers and instructors to administer the rubric in similar social and environmental settings if their goal is to assess growth in individual students (Phonethibsavads, Bender, & Pepler, 2019). Student performances should only be scored according to the impartial criteria described in the rubric. Appendix A shows the criteria for scoring in each category for the specific warm-up activity used during the rubric reliability testing. Teachers and other users of this rubric can adapt the specific criteria according to the demands of the activity, and it should state those criteria explicitly in the rubric. By stating criteria explicitly, the rubric provides students with actionable items to improve upon, and this concrete feedback may mitigate discouragement should students perform poorly (Geister et al., 2006).

3. Methods

In order to check the rubric's reliability, validity, and feasibility, we implemented a quasi-experimental design to explore pre-post changes in drama performance as measured by the rubric. The drama activity chosen for validation of this version of the rubric was a common warm-up exercise, the name-and-movement game. In it, a student says their name and performs a movement, and then the entire class imitates the name and movement together, before moving on to the next student. We videotaped the student performances in order to allow for flexibility in the scoring method (i.e., utilizing raters who could not be present for the performances).

3.1. Intervention context of the study

Students in schools participating in the Inner-City Arts (ICA) Learning and Achieving through the Arts (LATA) program received drama instruction over the course of a school year from professional teaching artists. The LATA Drama program involved going to the ICA campus two days a week for 14 weeks for intensive 3-hour drama classes during the school day. All treatment classes received instruction from the same teaching artist. During the 14-week term, the instructor taught drama concepts such as story structure, characterization, conveying emotion, and improvisation; trained students in the processes of rehearsing and performing; and focused on the relationship between

actors and audience. Students in the control classes did not receive treatment. However, the teachers of those classes arranged to receive treatment in future iterations of the LATA program.

3.2. Sample selection

Participants were drawn from four public elementary schools in the Los Angeles Unified School District (LAUSD). Three fourth-grade classrooms were chosen as the treatment group from two schools participating in the drama intervention (i.e., classrooms A, B, and C), and three fourth-grade classrooms from two schools that did not participate in the drama intervention were chosen for the control group (classrooms D, E, and F). More than 30 criteria were used to ensure comparability between treatment and control schools, including similar baseline standardized test scores, attendance rates, reclassification rates, parent participation, school suspension, safety, student demographics, English Language Learner progress rates, total enrolment, quality of facilities, and indicators of teacher quality (Pepler, Catterall, & Bender, 2015). In coordination with the district, we only had access to aggregates and not to the student level data, which is reflected in the type of analysis we were able to perform. In total, 97 students participated in the treatment group (41 boys and 56 girls), and 80 in the control group (39 boys and 41 girls). All students were ages 9–10, and four of the six classes were predominantly Hispanic (51–84%), while the other two had predominantly an Asian population (10–41%). In total, there was a small population belonging to other groups (2–10%). The aggregate majority of students were of lower socioeconomic status (69–100%). The difference in numbers is primarily due to the larger class sizes prevalent in the treatment schools. All students were assessed using the rubric twice (once at the start and again at the end of the study) on their performance of the same warm-up exercise.

3.3. Data collection

Observation data for the treatment classrooms were collected by videotaping each class's students completing the drama activity at the beginning and end of the 14-week LATA drama programme during the 2012 school year. Observation data for the control classrooms were collected in the same way but in the subsequent school year (2013). Data collection for the treatment classrooms took place in a black box theatre on the ICA campus, while for the control classrooms, it took place at the students' school in their everyday classroom. At each assessment point, the instructor gave the same instructions and modeled the activity. The researchers observed minimal differences in the way that the instructions and modeling were conducted among classrooms and assessment points. A professional videographer recorded the activity, focusing on each student and the students on either side of the focal student performing the activity. The videographer stayed in the center of the circle and rotated as each child took their turn performing their name and movement. The videographer remained as unobtrusive as possible. Students were aware that they were being filmed, but they were not made aware of the categories that they would be scored on beforehand because the rubric was still in development during data collection. In the data collection process, we followed IRB protocol by obtaining informed consent from teachers and parents and assent from students.

3.4. Limitations

Students were not made aware of the categories that they would be scored on beforehand because the rubric was still in development during data collection. We also only have data from two performances for each student, so that limited the students' ability to demonstrate their skills. The rubric also was designed to be specific to a single activity in a particular setting, and it was intended to assess learning between the beginning and end of the intervention. Evaluators and educators should

feel free to modify it for other activities and implement the assessment intermittently throughout the intervention, should they need more time points to assess growth. The statistics we calculated for reliability and validity may not hold under different activity circumstances.

4. Results

4.1. Establishing inter-rater reliability

We investigated inter-rater reliability to gauge the rubric's stability under varying circumstances. In order to establish the reliability of the rubric via acceptable inter-rater agreement, two external raters scored 29% of the data, and their scores were systematically compared. Prior to scoring this portion of the data, the raters underwent a training process which involved watching the video footage of a few students at a time, scoring them independently, and then comparing the scores, discussing and resolving discrepancies. Raters continued to engage in this process until they were able to score five students in a row in the same way. The movement and verbal response were evaluated together as one performance. Training required a total of 18 cases (students). After the reliability training was complete, the raters then scored 85 cases from the pre- and post-assessments of control classrooms D and E, representing 29% of the total data (total $N = 292$, 148 pre and 144 post). Because the scoring scale is ordinal rather than interval, the gamma statistic was used to calculate inter-rater reliability (Ruiz & Hüllermeier, 2012). The gamma statistics were acceptable ($G > 0.7$) for all scoring categories: Diction and volume (0.912), Movement and gesture (0.834), Group coordination (0.881), and Stage presence (0.819). When all categories were combined into the total score, the resulting gamma between the two raters was 0.725. Future work will consider test-retest reliability and internal consistency reliability, which we further discuss in the last section.

4.2. Validity

We focus on convergent and discriminant validity because of our small sample size, and, in practice, drama education relies on the subjective appraisals of experts (Oreck et al., 2003). By establishing validity with respect to expert opinion, the rubric may provide inferences that are consistent with standard practice. A convergent validity test was conducted in order to determine correlations between ratings on our rubric and an external expert's ratings on both a related (convergent) and unrelated (discriminant) construct, which in this case was creativity.

4.2.1. Convergent validity

To establish convergent validity, an expert in drama education was asked to rate a random subsample of 20 student performances (videos) for "technical proficiency." The expert rater was a professor in drama education, with a doctorate in dramaturgy and over 30 years of performance and teaching experience. This was a simple 1–5 Likert scale rating, with 1 representing low proficiency and 5 representing high proficiency, with no additional guidelines provided. The expert utilized his own subjective opinions. We then compared the non-expert raters' ratings with the drama expert's ratings to evaluate the level of consensus; if the non-expert raters were able to use the rubric to make the same judgments as the expert, then that consensus would provide evidence for convergent validity (Cable & DeRue, 2002). Due to the ordinal nature of the scoring scale, we used Spearman's rank correlation coefficient. A Spearman's correlation between the expert's ratings and the non-expert raters' ratings of the same students yielded an r of 0.594. This is a moderate correlation (Rea & Parker, 2014), showing that scores from the rubric mostly converge with scores from an expert. This evidence of convergent validity is moderate, but future work should improve on convergent validity. It is possible that the lower-than-ideal convergent validity here is due to the small sample size and a lack of

variety in performance levels among the students in the sample.

4.2.2. Discriminant validity

The expert rater also rated a random subsample of 120 performances for creativity in order to test whether the rubric correlates highly with measures of creativity. This rating was done using a 1–5 Likert scale with 1 representing 'low' creativity and 5 representing 'high' creativity. We operationalized creativity in a sociocultural framework, so we looked at the contributions individuals make to their community (e.g., individual performances in the context of the drama class). Creativity is inherently subjective, so the new contribution is only "creative" if the target audience appraises it as such (Phonethibsavads, Bender, & Pepler, 2019). Thus, the expert rater was prompted to use his own subjective criteria for judging creativity as required by Amabile's Consensual Assessment Technique for evaluating creativity (Amabile, 1982). Since creativity is a different construct from drama and performance skills, it was expected that his creativity scores would not correlate highly with the scores from the drama performance rubric. A Spearman's correlation between the expert's ratings of creativity and the non-expert raters' total scores of these performances when rated with the rubric yielded an r of 0.558, a moderate correlation. This indicates that the rubric scores may not be entirely independent of creativity, which may be due to aspects of the performed activity involving improvisation, a performance skill conceptually related to creativity.

4.3. Evaluating classroom differences

With the ordinal nature of the scoring scale, normality can not be assumed. Therefore, to evaluate classroom differences we used non-parametric tests of equivalence of median values (also considering the unequal sample size of the treatment and control groups). First, we conducted two-sample Wilcoxon rank-sum tests (Wilcoxon, 1947) for the pre-assessment of the treatment and control groups in order to check that both groups started out with an equivalent baseline. Then we conducted Wilcoxon matched-pairs signed-rank tests (Wilcoxon, 1947) between pre- and post-assessment for the control group to test whether there were any significant differences when no treatment occurred. We conducted the same test for the treatment group to test whether the treatment had a significant effect. Variances between the treatment and control usually varied greatly, so equal variances were not assumed. The alpha level was 0.05. Median and range values for all treatment and control scores are displayed in Table 2, as these two descriptive statistics are more appropriate for ordinal data.

No significant differences were found between the treatment and control groups' pre-test scores in the Diction/Volume, Movement/Gesture, or Group Coordination categories. However, there was a significant difference between the treatment and control groups in the Stage Presence category, with the treatment group starting off with significantly lower scores than the control group ($z = 2.57$, $p < 0.05$). The total scores of the treatment and control groups in the pre-test were also not significantly different from each other, which indicates that they were relatively equivalent before the treatment occurred.

As expected, in the control condition, no significant differences were found between the pre-test and post-test scores for any of the categories. There was also no significant difference between the pre-test and post-test of the control group's total scores. This indicates that scores as a group remained similar on average to each other between pre and post, as one would expect for classrooms that received no intervention. The treatment classrooms also performed largely as expected with significant gains between pre and post in all scoring categories. Aside from the first category, Diction/Volume, gains from pre-test to post-test were statistically significant at $p < 0.001$ (see Table 3). The total scores for the pre-test and post-test were also significantly different from each other, with the post-test scores higher than the pre-test scores on average ($z = 5.82$, $p < 0.001$). Thus, the intervention appeared to have a positive effect on students' performances, which the instrument was able to

Table 2
Median and Range Values for Treatment and Control Pre-test and Post-test Score Results Scored by LATA Drama Performance Rubric.

Treatment	Pre-test			Post-test			
	N	Median	Range	N	Median	Range	
Diction/Volume	96	3	1–5	93	3	1–5	
Movement/ Gesture* **	96	3	1–5	93	4	1–5	
Group	96	3	1–5	93	4	1–5	
Coordination* **							
Stage Presence* **	96	3	1–4	93	4	2–5	
Total score* **	96	12	6–18	93	14	9–19	
Control	Pre-test			Post-test			
	N	Median	Range	N	Median	Range	
	Diction/Volume	75	3	2–5	72	3	2–5
	Movement/ Gesture	75	3	2–5	72	3	2–5
	Group	75	3	1–5	71	3	1–5
	Coordination						
	Stage Presence	75	3	2–4	72	3	2–4
Total score	75	12	9–18	72	12.5	9–18	

An asterisk (***) indicates a significant ($p < 0.001$) difference between pre and post.

Table 3
Wilcoxon Matched-Pairs Signed-Rank Test for Treatment and Control between Post- and Pre-Assessment.

	Treatment			Control		
	N	z	p	N	z	p
Diction/Volume	92	1.72	0.09	67	-0.73	0.467
Movement/ Gesture	92	3.46	0.0005 ***	67	1.49	0.136
Group Coordination	92	4.60	0.0000 ***	67	1.64	0.1
Stage Presence	92	6.29	0.0000 ***	67	0.84	0.404
Total score	92	5.82	0.0000 ***	67	1.53	0.126

An asterisk (***) indicates a significant ($p < 0.001$) difference between pre and post.

detect.

We also performed an analysis to check whether the instrument could distinguish between students who had received the intervention and those who had not. To this end, we performed two-sample Wilcoxon rank-sum tests, testing the difference between the treatment and control change scores (post-test score – pre-test score) for each rubric category and for the total score. The results showed that there were significant differences in the change score between the treatment group and the control group in all categories except the Diction/Volume category ($z = -1.72, p = 0.09$). The remaining three categories were all statistically significant, as follows: Movement/Gesture ($z = -2.40, p = 0.02$); Group Coordination ($z = -1.98, p = 0.05$); Stage Presence ($z = -4.61, p < 0.001$). Changes in Total Scores were statistically significantly different between treatment and control classrooms ($z = -4.09, p < 0.001$). Category-specific analysis indicated that the Stage Presence change score contributed the most to the difference between treatment and control groups’ change in Total Scores.

5. Discussion, limitations, and recommendations

This study presented an initial foray into addressing the need for early elementary standards-based drama assessments that have been tested for reliability and validity and that evaluators and other stakeholders in the arts might find useful. By outlining the process of creating the instrument, we further aimed to shed some insight on the challenges of creating such a rubric, and on the strengths and limitations of the product, with an eye toward next steps for researchers and practitioners. For the most part, the above results demonstrated that the tested rubric is both reliable and valid for use in an early elementary school drama context: its content was validated by consulting with drama teaching artists and by aligning to standards; it showed high inter-rater reliability; it can distinguish between classrooms that have received drama instruction and those that have not. Further work is needed on

convergent and discriminant validity.

A limitation of the quasi-experimental design is that we did not employ other measures of reliability such as test-retest reliability and internal consistency. Test-retest reliability was not appropriate for the treatment group since we hypothesized that scores would change over time. Even though we collected data for the control group at two different time points, there was still a significant amount of time between the two (14 weeks), suggesting that there might be some changes, albeit not statistically significant. Future research could test the control group at another time point, closer to the first to establish test-retest reliability and examine the internal consistency of the rubric, as another indicator of its reliability.

A strength of the assessment is that it took place in the context of a short warm-up activity in a low-stakes environment. As such, it holds the potential to be an unobtrusive part of ongoing evaluation efforts of drama learning toward providing evidence to funders. We hope that educators and other stakeholders might also find the instrument useful. As the activity captures learning over time, the rubric could be used at different time points to evaluate student development over time across the select skills, measured by the instrument. In this study for the purposes of evaluation, we recorded student performances and performed extensive inter-rater reliability testing in order to test the instrument. Classroom drama teachers could use the rubric in real time as students engage in the activity to estimate scores. Although the experience of drama might never be captured fully in numbers, the scores can serve as valuable discussion points to engage students in authentic self-assessment (Oreck et al., 2003; Somers, 2005). The feedback on the rubric could thus provide actionable steps for students on how to improve, which could mitigate feelings of discouragement. Still, educators must be cautious about coming to summative conclusions from only a single administration of the assessment and perhaps the instrument is best used for formative purposes.

We must caution that a single exercise like the name-and-movement activity is not necessarily indicative of a student’s ‘true ability’ in drama. For one, there are many reasons why a student capable of performing better may not have performed to the best of their ability during administration of the assessment. Teachers should be cognizant of social and environmental factors at the time of assessment because some students would likely be inhibited by performance anxieties in this work. We must also caution that due to the nature of the name-and-movement activity, the assessment rubric is only set to evaluate a select few drama skills while missing key elements such as a focus on character development, partner scene work, etc. (Phonethibavads, Bender, & Pepler, 2019). Additional research is needed to address the inclusion of other foundational skills in a revision of the present assessment rubric, or in perhaps the creation of a more universal tool that is better suited for a range of activities. However, in turn this raises the question as to how

many assessment rubrics and tools are really necessary, wherein lies one of the challenges we faced in constructing this assessment tool – finding a balance in the trade-off between utility, practicality, and universality. For now, this initial effort to create a research-based assessment toward the support of program evaluators and other interested stakeholders shows promise in its reliability, validity, and ability to assess drama learning over time.

We must also remind educators, researchers, and evaluators that the rubric was developed as a distillation of concrete observations, so such a rubric should not take precedence over one’s own aesthetic sense. Historically, drama performances have always been assessed by the intuitions and opinions of trained experts, and the rubric is intended to coordinate judgments of drama experts and non-experts so that it may be possible to organize consistent educational interventions. Especially for elementary teachers, who are likely to teach multiple subjects rather than being certified in drama or theater specifically, this type of rubric can allow general education teachers to reliably assess students’ drama skills and understanding. To streamline the process for non-expert raters, perhaps a revision of the rubric scale is needed toward improving its reliability. At present, raters score performances on a 1–5 Likert type scale, but it could be that such nuance is not necessary. A 3-point scale assessing the mastery of a given skill from emerging through developing and to achieved, as in the case of Susman-Stillman and colleagues’ PTAR (2018) might be more fitting and easier to use.

Assessment rubrics that align with multiple states’ drama standards, can save teachers’ time and effort, allowing them to focus primarily on instruction instead of designing their own assessments. Appendix B shows the rubric’s alignment with the NCAS. It also suggests revisions to the rubric’s scoring category names toward better alignment with the NCAS. As is, the rubric currently captures only two out of the four domains put forth by the NCAS, creating and performing, as a result of the activity in focus. We recommend that if the rubric is revised to be used for an activity other than the name-and-movement game and/or to address the other two domains, responding and connecting, then revisions of the rubric should start with looking at the anchor standards that are most appropriate for the context of the activity. In some cases, existing categories could be adapted to reflect the new activity. For instance, if the new activity does not involve imitation of movement the way the name-and-movement exercise does, then some other action indicating respect for the group activity should be substituted. In other cases, including additional categories might be needed to capture

standards that have not yet been addressed in the present rubric. If the new activity involves aspects of reflection or aspects grounded in student prior experience, adding categories that fall under the remaining two NCAS domains (i.e., responding and connecting) might be needed.

Ultimately, the development of this scoring instrument is an important first step towards reliable, valid, practical, and standards-based assessment of select skills in drama performance that can support evaluation efforts. This helps to fill a gap in drama assessment, so external assessors, educators and other stakeholders can now better gauge the effects of drama instruction on drama learning over time or relate the learning of drama skills to the learning of other valued skills. The LATA Drama Performance Rubric is presented here as one way to help the field of drama education forge ahead in an American educational landscape increasingly dominated by standards-based content and by the need for evidence of learning.

Funding details

This work was funded by the US Department of Education’s Arts in Education Model Development and Dissemination program.

CRedit authorship contribution statement

Kylie Pepler: Conceptualization, Methodology, Funding acquisition, Supervision, Writing – original draft. **Sophia Bender:** Investigation, Formal analysis, Writing – original draft, Visualization, Validation. **Anthony Phonethibsavads:** Investigation, Writing – review & editing. **Nickolina Yankova:** Writing – review & editing. **Molly Stewart:** Formal analysis, Visualization, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Beth Tishler, Jan Kirsch, Kristy Messer, Inner-City Arts, Los Angeles for their support and feedback on this work.

Appendix A. Original LATA Drama Performance Rubric

Category	Coding	Criteria
Diction and volume	5	Pronunciation is exceptionally loud and clear
	4	Pronunciation is both loud and clear
	3	Pronunciation is either loud or clear, but not both
	2	Pronunciation can be heard, but not fully understood
	1	Pronunciation can be heard barely or not at all
Movement and Gesture	5	Motions are exceptionally original, expressive, and/or assertive
	4	
	3	Movement is moderately original, expressive, and/or assertive
	2	
	1	Movement is unoriginal, timid, unspecific, or lacking bold expression, or not performed
Teamwork/specificity of imitative movement	5	Student imitates motions with attention to specific details in gestures and posture and matches the model’s expressiveness and assertiveness
	4	Student imitates motion with attention to specific details in gesture and posture
	3	Student imitates motions recognizably, but misses some details
	2	Student performs an imitation, but it is quite different from the model
	1	Student’s imitation differs from the model, is half-hearted and cursory, and/or imitated movement is not performed
Stage presence/maintaining neutral position	5	Exhibits focus when both improvising and imitating. Is present in the moment and ready to act. Use of neutral position when not actively performing. Appears relaxed and responsive.
	4	Student occasionally deviates from neutral position or is slightly slow to act or slightly off-task.
	3	Student is focused and in neutral position about half of his/her time on camera
	2	Student lacks focus and/or maintenance of the neutral position during most of his/her time on camera.
	1	Student does not understand the concept of the neutral position, is slow to act, and requires much coaching; and/or does not perform the required activity.

Appendix B. Comparison of original and current standards and rubric categories

Original Rubric Category	Recommended Rubric Revisions	National Core Arts Standards (2014)
Diction and volume	Expression through sound and voice	Cr 3.1b (2): 'Use and adapt sounds and movements in a guided drama experience' and Pr4.1b (3): 'Investigate how movement and voice are incorporated into drama/theatre work.'
Movement and gesture	Expression through movement	Cr2b (K): 'With prompting and support, express original ideas in dramatic play or a guided drama experience'; also Cr 3.1b (2) and Pr4.1b (3).
Teamwork/specificity of imitative movement	Group coordination	Cr2b (7): 'Demonstrate mutual respect for self and others and their roles in preparing or devising drama/theatre work' and Pr5.1a (3): 'Participate in a variety of physical, vocal, and cognitive exercises that can be used in a group setting for drama/theatre work.'
Stage presence/maintaining neutral position	Behavioral expectations	When not actively performing, student listens to other members' performances and maintains focus and readiness to contribute to the activity appropriately, which may be planned or improvised.

References

- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013. <https://doi.org/10.1037/0022-3514.43.5.997>
- Baptiste, L. (2008). Managing Subjectivity in Arts Assessments. *Reconceptualising the Agenda for Education in the Caribbean*, 503.
- Barlow, W. D. (2020). 'We ur al aff tae th'big schuil'—pupils' and teachers' views and experiences on using Drama Conventions to support primary-secondary transition. *Education 3-13*, 48(8), 893–908.
- Bolton, G. (2007). A history of drama education: A search for substance. *International Handbook of Research in Arts Education* (pp. 45–66). Dordrecht: Springer., https://doi.org/10.1007/978-1-4020-3052-9_4
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24(1), 61–100. <https://doi.org/10.3102/0091732x.024001061>
- Brown, V. (2017). Drama as a valuable learning medium in early childhood. *Arts Education Policy Review*, 118(3), 164–171. <https://doi.org/10.1080/10632913.2016.1244780>
- Cable, D. M., & DeRue, D. S. (2002). The convergent and discriminant validity of subjective fit perceptions. *Journal of Applied Psychology*, 87(5), 875. <https://doi.org/10.1037/0021-9010.87.5.875>
- Chen, F., Lui, A. M., Andrade, H., Valle, C., & Mir, H. (2017). Criteria-referenced formative assessment in the arts. *Educational Assessment, Evaluation and Accountability*, 29(3), 297–314. <https://doi.org/10.1007/s11092-017-9259-z>
- Deasy, R. J., 2002, Critical links: Learning in the arts and student academic and social development. Arts Education Partnership.
- DeBettignies, B. H., & Goldstein, T. R. (2020). Improvisational theater classes improve self-concept. *Psychology of Aesthetics, Creativity, and the Arts*, 14(4), 451.
- Fleming, M., Merrell, C., & Tymms, P. (2004). The impact of drama on pupils' language, mathematics, and attitude in two primary schools. *Research in Drama Education*, 9(2), 177–197. <https://doi.org/10.1080/1356978042000255067>
- Foster, E. M., & Marcus Jenkins, J. V. (2017). Does participation in music and performing arts influence child development. *American Educational Research Journal*, 54(3), 399–443. <https://doi.org/10.3102/0002831217701830>
- Gallagher, K., & Service, I. (2010). Applied theatre at the heart of educational reform: An impact and sustainability analysis. *Research in Drama Education: The Journal of Applied Theatre and Performance*, 15(2), 235–253. <https://doi.org/10.1080/13569781003700144>
- Geister, S., Konrad, U., & Hertel, G. (2006). Effects of process feedback on motivation, satisfaction, and performance in virtual teams. *Small Group Research*, 37(5), 459–489. <https://doi.org/10.1177/1046496406292337>
- Greenfader, C. M., Brouillette, L., & Farkas, G. (2015). Effect of a performing arts program on the oral language skills of young English learners. *Reading Research Quarterly*, 50(2), 185–203. <https://doi.org/10.1002/rrq.90>
- Hagen, U., & Frankel, H. (1973). *Respect for Acting*. John Wiley & Sons.,
- Hardiman, M., Rinne, L., & Yarmolinskaya, J. (2014). The effects of arts integration on long-term retention of academic content. *Mind, Brain, and Education*, 8(3), 144–148. <https://doi.org/10.1111/mbe.12053>
- Van de Water, M., McAvoy, M., & Hunt, K. (2015). *Drama and education: Performance methodologies for teaching and learning*. Routledge., <https://doi.org/10.4324/9781315756028>
- Jacobs, R. (2016). Challenges of drama performance assessment. *Drama Research*, 7(1), 2–19.
- Jacobs, R. (2022). Assessment in drama education. *The Routledge Companion to Drama in Education* (pp. 137–150). Routledge.,
- Johnstone, K., & Wardle, I. (2012). *Impro: Improvisation and the Theatre*. Routledge.,
- Kelner, L. B., & Flynn, R. M. (2006). *A Dramatic Approach to Reading Comprehension: Strategies and Activities for Classroom Teachers*. Heinemann.,
- Kisida, B., Goodwin, L., & Bowen, D. H. (2020). Teaching history through theater: The effects of arts integration on students' knowledge and attitudes. *AERA Open*, 6(1), 1–11. <https://doi.org/10.1177/2332858420902712>
- Korkut, P. (2018). The construction and pilot application of a scoring rubric for creative drama lesson planning. *Research in Drama Education: The Journal of Applied Theatre and Performance*, 23(1), 114–125. <https://doi.org/10.1080/13569783.2017.1396211>
- Korošec, H., & Zorec, M. B. (2020). The impact of creative drama activities on aggressive behaviour of preschool children. *Research in Education*, 108(1), 62–79.
- Li, X., Kenzy, P., Underwood, L., & Severson, L. (2015). Dramatic impact of action research of arts-based teaching on at-risk students. *Educational Action Research*, 23(4), 567–580. <https://doi.org/10.1080/09650792.2015.1042983>
- Lin, M. C. (2013). The development of a performance assessment with performing arts teachers in Taiwan—from national policy to classroom practice. *Research in Drama Education: The Journal of Applied Theatre and Performance*, 18(3), 296–312. <https://doi.org/10.1080/13569783.2013.810926>
- Linklater, K. (2006). *Freeing the Natural Voice: Imagery And Art in the Practice of Voice and Language*. Nick Hern Books.,
- Love, A. J. (2010). Understanding approaches to evaluation.
- Ludwig, M. J., Boyle, A., & Lindsay, J. (2017). Review of evidence: Arts integration research through the lens of the Every Student Succeeds Act. *American Institutes for Research*.
- Mages, W. K. (2018). Does theatre-in-education promote early childhood development?: The effect of drama on language, perspective-taking, and imagination. *Early Childhood Research Quarterly*, 45, 224–237.
- Mathison, S. (2010). The purpose of educational evaluation.
- Meyer, M. J. (2016). Arts-Inspired Performance Assessment Considerations for Educational Leaders. *Leadership of Assessment, Inclusion, and Learning* (pp. 111–140). Cham: Springer., https://doi.org/10.1007/978-3-319-23347-5_5
- National Coalition for Core Arts Standards, 2014, National Core Arts Standards. (<http://www.nationalartsstandards.org/>).
- National Coalition for Core Arts Standards, 2016, Theatre model cornerstone assessments. (<https://www.nationalartsstandards.org/mca/theatre>).
- National Coalition for Core Arts Standards, 2019, The status of arts standards revision in the United States since 2014. (https://www.nationalartsstandards.org/sites/default/files/NCAS-StateReport_2019_digital-FINAL.pdf).
- National Coalition for Core Arts Standards, 2020, Arts education for America's students: A shared endeavor. (<http://www.nationalartsstandards.org/sites/default/files/A%20Shared%20Endeavor%209.15.14.pdf>).
- Omasta, M., Murray, B., McAvoy, M., & Chappell, D. (2021). Assessment in elementary-level drama education: Teachers' conceptualizations and practices. *Arts Education Policy Review*, 122(4), 265–279.
- Oreck, B. A., Owen, S. V., & Baum, S. M. (2003). Validity, reliability, and equity issues in an observational talent assessment process in the performing arts. *Journal for the Education of the Gifted*, 27(1), 62–94. <https://doi.org/10.1177/016235320302700105>
- Peppler, K., & Catterall, J. S. (2006). *Year two findings on the arts learning of children enrolled in the LA's BEST after school arts program (Deliverable to the LA's BEST After-School Arts Program)*. Los Angeles, CA: University of California.
- Peppler, K., Catterall, J. S., & Bender, S. (2015). *Learning and achieving through the arts: A collaborative project of inner-city arts and Los Angeles Unified School District 4 (Deliverable to the U.S. Bloomington: Department of Education)*.
- Peppler, K., Catterall, J., & Feilen, K. (2009). *Arts in the middle: A collaborative project of Inner-City Arts and Los Angeles Unified School District 4 (Deliverable to the U.S. Bloomington: Department of Education)*.
- Phonethibavads, A., Bender, S., & Peppler, K. (2019). Utilizing the consensual assessment technique to compare creativity in drama spaces. *Creativity Theories—Research-Applications*, 6(1), 4–19.
- Ragot, L. (2011). Assessing student learning by way of drama and visual art: A semiotic mix in a course on cognitive development. *Education as Change*, 15(sup1), S63–S78. <https://doi.org/10.1080/16823206.2011.643625>
- Rea, L. M., & Parker, R. A. (2014). *Designing and Conducting Survey Research: A Comprehensive Guide*. John Wiley & Sons.,
- Rose, D. S., Parks, M., Androes, K., & McMahon, S. D. (2000). Imagery-based learning: Improving elementary students' reading comprehension with drama techniques. *The Journal of Educational Research*, 94(1), 55–63. <https://doi.org/10.1080/00220670009598742>
- Schwartz, D. L., & Arena, D. (2013). *Measuring What Matters Most: Choice-based Assessments for the Digital Age*. The MIT Press., <https://doi.org/10.7551/mitpress/9430.001.000>

- Somers, J. (2005). Measuring the shadow or knowing the bird: Evaluation and assessment of drama in education. *Evaluating Creativity* (pp. 116–137). Routledge.
- Susman-Stillman, A., Englund, M., Webb, C., & Grenell, A. (2018). Reliability and validity of a measure of preschool children's theatre arts skills: The Preschool Theatre Arts Rubric. *Early Childhood Research Quarterly*, 45, 249–262. <https://doi.org/10.1016/j.ecresq.2017.12.001>
- Walker, E. M., McFadden, L. B., Tabone, C., & Finkelstein, M. (2011). Contribution of drama-based strategies. *Youth Theatre Journal*, 25(1), 3–15. <https://doi.org/10.1080/08929092.2011.569471>
- Walker, E., Tabone, C., & Weltsek, G. (2011). When achievement data meet drama and arts integration. *Language Arts*, 88(5), 365.
- Weltsek, G. J., Duffy, P. B., & Carney, C. L. (2014). The local and global state of theater education research and policy. *Arts Education Policy Review*, 115(3), 63–71. <https://doi.org/10.1080/10632913.2014.913968>
- Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, 3(3), 119–122.
- Winner, E., & Hetland, L. (2004). Cognitive transfer from arts education to non arts outcomes: Research evidence and policy implications. *Handbook of Research and Policy in Art Education* (pp. 143–170). Routledge.