

Utilizing the Consensual Assessment Technique to Compare Creativity in Drama Spaces*

Anthony Phonethibsavads

Indiana University, USA

E-mail address: anphonet@indiana.edu

Kylie Peppler

University of California, Irvine, USA

E-mail address: kpeppler@uci.edu

Sophia Bender

Indiana University, USA

E-mail address: sobender@indiana.edu

ARTICLE INFO

Keywords:

Creativity

Drama

Consensual assessment

Education

Article history:

Received 22 August 2018

Received in revised form 03 February 2019

Accepted 01 April 2019

ISSN: 2354-0036

DOI: 10.1515/ctra-2019-0001

ABSTRACT

Objective: This study examines the validity of Amabile's (1982) consensual assessment technique in measuring creativity in a warm-up activity in fourth-grade drama classrooms and compares the scores between warm-ups occurring in a blackbox theater setting (experimental) vs. a traditional classroom (control). **Method:** Four professional actors viewed 60 clips of children's drama warm-ups and scored for creativity, using a 5-point scale. After establishing sufficient inter-rater reliability (IRR), we used the average scores of the raters to compare creativity between the experimental and control groups. **Results:** The raters demonstrated high agreement, with a coefficient alpha estimate of .819. An independent samples t-test between the experimental and control groups was significant at $p < .001$, with the experimental group receiving higher scores. **Conclusions:** The results suggested that creativity was significantly higher in the experimental group, and the context correlated with creativity, despite neither group having yet received drama instruction at that time. This paper presents discussions about validity, opinions of the raters, possible implications for the activity itself, and possible effect of setting on creativity.

INTRODUCTION

When people observe creativity in everyday life, we typically observe the creativity of finished products (Csikszentmihalyi, 1988; Plucker, Beghetto, & Dow, 2004), and the social and environmental conditions surrounding the process is largely irrelevant (Nachmanovitch, 1990). Society values creativity that leads to the development of newer

* This study was made possible by funding from the National Science Foundation and by the late Dr. James Catterall, whose contributions helped greatly promote arts education for better creativity and closing the achievement gap in schools.

and better products (Csikszentmihalyi, 1988), so the consensual assessment technique (CAT; Amabile 1982) has become a standard because it utilizes a panel of experts to appraise the creativity of final products, rather than the abilities of individuals. However, from the perspective of educators, it behooves us to broaden assessment practices to include person-in-situ interactions because the environment sends powerful cues, both conscious and unconscious, that influence behaviours and decisions that lead to the creation of a final product (Barab & Plucker, 2002; Beghetto & Kaufman, 2014). Although the emphasis on product creativity contributed to the development of the consensual assessment technique (CAT; Amabile, 1982), we argue that this instrument can be applied to the assessment of drama warm-ups and supports comparing creativity between different contexts (Podlozny, 2000).

This study explores two questions: (1) To what extent is the CAT valid for measuring creativity in drama? And (2) Can the CAT distinguish between creativity under different social and environmental conditions? Our findings suggest that the CAT yielded valid inferences for this context because our raters observed drama warm-ups and reported strong inter-rater reliability (IRR; $\alpha = .819$). Additionally, an independent samples *t*-test between the *experimental* and *control* groups was significant at $p < .001$ with the *experimental* group receiving higher scores, which suggests a strong correlation between environment and creativity. Validity issues will be discussed.

BACKGROUND

Prior research on contexts suggests that social and environmental factors influence creativity: Sawyer's (2003) work on improvisation suggests that creativity flourishes in social settings where individuals affirm each other's contributions and elaborate on them without judgment. Further, Beghetto and Kaufman's (2014) work on classrooms suggests that the environment sends cues, both conscious and unconscious, to students that can trigger insights that would facilitate the creation of novel and valuable products. If we accept these premises, then we cannot think about creativity as a stable trait possessed by individuals, but instead, as a conflux of dynamic and evolving transactions between individuals and contexts (Barab & Plucker, 2002). By examining the dialectical interactions between persons and contexts, we can understand how individuals appear to be creative if their contributions are culturally sanctioned as novel and valuable (Plucker, Beghetto, & Dow, 2004).

If we are examining creativity as interactions between persons and environments, then we must not only understand what constitutes creativity, but we must also investigate where creativity happens (Csikszentmihalyi, 1988). According to Csikszentmihalyi's systems model, creativity is a transaction between *persons*, *domains* and *fields*: persons

appropriate knowledge from a domain and create new products based on trends and aesthetics, and these products are submitted to a field. The field determines whether these products should be preserved or rejected, and usually, products are preserved if they advance the goals of the field. At the level of a drama class, one student's utterance or performance can inspire other students, who would emulate or elaborate upon the previous performance (Sawyer, 2003).

Because we are interested in examining creativity at the level of interactions rather than products, our study examines a warm-up exercise rather than a showcase. By examining a warm-up, we can minimize external pressures that audiences may present and better focus on the learning of individuals acting with their environments (Barab & Plucker, 2002; Oreck, Owen, & Baum, 2003). Additionally, because raters can distinguish between process vs. product creativity (Hennessey, 1994), the warm-up may provide a more valid representation of creativity in the learning process than a showcase would. Furthermore, because students in drama classes are trained to react and adapt to unexpected changes in their environment, performances are easily influenced by changes in time and space—especially in unscripted settings (Sawyer, 2000; Scruggs & Gellman, 2007). Therefore, a valid representation of process should include the possibility of spontaneous changes (Nachmanovitch, 1990; Oreck, Owen, & Baum, 2003). Thus, if the contexts of different performances remain consistent over time while the performances themselves change in quality, then we can claim that the reactions or interactions that occur within one context are more conducive toward creativity than those that occurred in another context where no qualitative improvement occurred (Beghetto & Kaufman, 2014; Sawyer, 2003).

The Consensual Assessment Technique (CAT)

Much like real-world evaluation of creativity (Csikszentmihalyi, 1988), the CAT depends on gaining approval from a community before going further: it utilizes a panel of experts to appraise the creativity of products and raters assign scores based on their own perceptions (Amabile, 1982). This technique differs from cognitive assessments, which focus on prompting individuals to complete novel tasks (e.g., use this squiggle to draw a picture, write a story entitled "The Flying Monkey," etc; Renzulli, 1978; Spicker, 1992; Torrance, 1972) and by examining responses, we gain secondary insight into the mental processes of the individual. Instead of focusing on the underlying processes, the CAT prompts judges to use their own subjective perceptions to assign numeric scores to products (Amabile, 1982); essentially, a response is creative if it is recognized by society as such (Csikszentmihalyi, 1988; Plucker, Beghetto, & Dow, 2004), so creativity is inherently sub-

jective. Although the cognitive mechanisms are not addressed, consensual assessment has proven valid for measuring differences in creativity between products created under different conditions (e.g., Amabile, Hennessey, & Grossman, 1986; Hennessey, 1994).

Although the CAT is a standard measure in creativity assessment, critics suspect that a subjective measure of creativity might yield variable scores (Baer & McKool, 2009). However, raters usually demonstrate considerable overlapping judgments if they belong to the same field (Advares-Yorno, Postmes, & Haslam, 2006; Amabile, 1982). The CAT is considered reliable if the inter-rater reliability (IRR) shows an alpha higher than .70 (Baer & McKool, 2009), which has been demonstrated repeatedly in different domains, such as poetry, music composition and painting (e.g., Baer, Kaufman, & Gentile, 2004; Hickey, 2001; Kaufman, Baer, Cole, & Sexton, 2008).

Not only has the CAT been valid for assessing creativity of products (i.e., finished artifacts that are presented to an audience), but also for processes leading to the completion of products (i.e., the preparation, decision-making and learning that was necessary before making an artifact presentable). For example, in Hennessey's (1994) study involving graphic design, raters were reliably able to assess the creativity of graphic design as both products and processes, distinguish between product and process in this context (i.e., graphic design) and their assessments of both products and processes highly correlated with their product scores. Because of the high correlation, it is possible to utilize the CAT to evaluate process, such as what we examined in the drama classrooms. However, the creative process in Hennessey's study was enacted by a computer program, so it was unaffected by presence of judges or changes in the environment and, arguably, not wholly relevant to drama performance. In drama, however, judges observe individuals interacting with other people and their environments, and the presence of judges would most likely alter the creativity of students in contexts (Barab & Plucker, 2002; Issacharoff, 1981; Llewellyn-Jones, 2002; Sawyer, 2000).

Why Drama Education?

Drama education is a longtime ally to creativity because role playing exercises help students take different perspectives and more deeply engage with texts in literature classes or scientific concepts in biology classrooms (Johnson, 2004; Podlozny, 2000; Walker, Tabone, & Weltsek, 2011). While activities teach students performance skills that can be used on stage, drama differs from theatre because the former emphasizes student learning (e.g., performing without an audience) while the latter emphasizes the production of shows (e.g., performing with an audience; Podlozny, 2000; Spolin, 1999; Weltsek, 2005). By examining performance exercises in a drama classroom, we witness interactions be-

tween individuals and contexts in real-time and by utilizing the CAT to appraise warm-ups, we opened possibilities for future examination of the interdependence of context and process (Nachmanovitch, 1990; Neelands, 2004; Spolin, 1999).

METHODS

Design

This study was part of the Learning and Achieving through the Arts (LATA) project, which was a four-year study funded by the United States Department of Education and examined the effect of arts-integrated education on Los Angeles-based elementary schools, using an experimental vs. control comparison (Author citation, 2015). An agreement was made with participating schools that classrooms serving as control groups would receive arts-integrated instruction in future years and vice-versa. In both conditions, students were given the same directions under the same instructor and raters were not informed about these conditions.

Participants

Students

A total of six classrooms (N=180) from Los Angeles-based schools participated in the study: three of the six classrooms that participated were taken to a blackbox theatre, with the knowledge that they would be given future drama instruction, and the other three served as a control group, remaining in their classrooms and never receiving drama instruction during the time of the study. All students were of the same age (9-10), but there were noticeable variations in ethnicity in each classroom: four of the six classes were predominantly Hispanic (51-84%), and 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%).

This study complied with ethical standards of human subjects research by obtaining permission for the study from Indiana University's Institutional Review Board, consent from teachers and parents, and assent from students.

Raters

In order to minimize the risk of raters misinterpreting creativity in this context, we assembled a panel of four professional actors, including three journeymen (Raters 1, 2, and 3) and one professor (Rater 4). Each journeyman had 2-5 years of experience performing at small venues and the professor possessed a doctorate in dramaturgy and more than 25 years of teaching and performance experience, including major productions. We began with three journeymen because it was the minimum number required to establish inter-rater reliability (Baer & McKool, 2009) and because more experienced experts were diffi-

cult to recruit, we included the professor to corroborate the responses from the first three raters. All raters followed directions and independently scored performances, and their scores were scaled to the ability level of the sample. Each rater was instructed to score the creativity of each child's own performance; although all children imitated their neighbours' actions as part of the warm-up, the imitations were not scored.

Procedure

Video Recording

We sought to present drama warm-ups so that raters could score them for creativity, but we had to ensure that our videos were an authentic representation of drama. Although showcases or auditions are the standard for assessing performance ability (Oreck, Owen, & Baum, 2003), we argue that warm-ups are valid for the following reasons: first, the activity involves some degree of "enactment"; students express themselves in pretend scenarios, usually through voice or movement (Podlozny, 2000). Second, there is structure to this "enactment"; students are given a "script" or "game" with a beginning, middle and end (Podlozny, 2000; Sawyer, 2000). And third, performances have the potential to be affected by changes in the environment (Barab & Plucker, 2002; Issacharoff, 1981; Llewellyn-Jones, 2002). Given these considerations, we chose the "Name and Movement" game for the activity, which is a warm-up where students stand in a circle, take turns saying their own name while performing a movement, and imitate their classmates' performances. Not only did this activity fit the criteria, but each student's performance was brief, making it manageable for raters to view in high volume. The video clips were embedded in an online survey programme, which was then distributed to four volunteer raters.

While we wanted to present a large number of short performances to our raters, we also had to be careful not to take these performances out of context. In order to preserve the context of performances, we recorded each whole-class exercise, then edited the footage into short segments that each showed the performance of the selected student as well as other students' interactions immediately preceding and following. Through the inclusion of these two performances, we were able to see whether or not students were affirming or elaborating on each other (Sawyer, 2003), which may have affected the ways that raters perceived creativity.

Assessment Tool

While it is usually ideal for raters to witness performances live, we acknowledged that the presence of judges could increase anxiety (LeBlanc, Jin, Obert, & Siivola, 1997) or incentivize students to perform boastfully (Scruggs & Gellman, 2007). Thus, we video recorded the student performances so that raters could view all whole-class recordings before

scoring individual performances. By viewing whole-class recordings, the rater acclimated to the social climate of each classroom, which may have influenced creativity perception. Additionally, the recordings allowed them to repeatedly review any performances.

Instructions at the top of each page prompted raters to score the “creativity” of each child’s choice of movement, not their imitations of their neighbours. The videos were divided across 12 pages, each with 5 hyperlinks to a video clip – all of which were presented in a different random order. Beneath each hyperlink was a 5-point Likert scale (1=low, 3=average, 5=high), where each rater independently provided scores, without deliberation with outside parties. Furthermore, they avoided unfair comparisons (e.g., comparing a child to a professional) by scoring for creativity relative to the sample. Additionally, there was a random sample of 20 performances at the end of the survey that prompted raters to score for “technical proficiency” rather than “creativity”; although Amabile (1982) advises raters scoring for overlapping constructs in addition to creativity, we did not prompt raters to score technique for all performances because it was time-consuming to view large numbers of video clips in succession, and we did not want to risk raters withdrawing participation due to fatigue.

Open-Ended Items

Although the CAT does not normally require raters to provide rationale for their scores, we found it necessary for two reasons: first, the CAT has not been applied to performance before, so the open-ended items provided complementary evidence that allowed us to make inferences in the event that the raters yielded insufficient inter-rater reliability (Johnson, Ongwubuzie, & Turner, 2007), and second, the inclusion of open-ended items provided an opportunity to probe raters’ perceptions (Isaken, Lauer, Ekvall, & Britz, 2001). It was important to examine raters’ perceptions because, in drama, the location of a performance is equally as important as the performance itself (Phelan, 2004), so by including open-ended responses, we could infer whether judges’ scores reflected actual differences in performance creativity, or whether performances merely appeared to be more creative in a different physical environment.

Following the survey, we asked the raters the following:

What did you look for in giving high scores (i.e., 4 or 5)?

What did you look for in giving low scores (i.e., 1 or 2)?

What did you look for in giving average scores (i.e., 3)?

FINDINGS

Question 1: To what extent is the CAT valid for measuring creativity in drama?

We first examined the creativity scores and established inter-rater reliability (IRR) by examining the alpha coefficient (Cronbach, 1951). Although many statisticians reject the

alpha coefficient as a valid measure of homogeneity (Cronbach & Shavelson, 2004; Schmitt, 1996), it remains a suitable estimate of IRR for assessing the creativity of products because it measures the correlations among raters' scores and more creative pieces should consistently receive higher scores while less creative pieces would consistently receive lower scores (e.g., Baer, Kaufman, & Gentile, 2004; Hickey, 2001; Kaufman, Baer, Cole, & Sexton, 2008). An alpha estimate ranging from .70 to .80 is considered to show strong agreement (Bear & McKool, 2009) and therefore, it was possible to proceed with a comparison of means.

We estimated an alpha of .819 for the four raters, which demonstrates strong agreement. While the raters gave similar scores to similarly creative performances, they did exhibit some variation in leniency; however, because the alpha coefficient measures the correlations between scores, the collective appraisal of this panel is reliable because each rater was consistent in their individual differences (Cronbach, 1951). Rater 1 was relatively "harsh," giving 10% of the performances a score of 1, 38.3% scores of 2, 26.7% scores of 3, 25% scores of 4, and not one score of 5; and Rater 2 was more "even," giving 33.3% a score of 2, 28.3% a score of 3, 31.7% scores of 4, and very few 1s and 5s. On the other hand, Raters 3 and 4 (professor) were similarly lenient: Rater 3 gave scores of 3 to 55% of the performances, 4s to 41%, very few 5s, and no 1s or 2s; similarly, Rater 4 gave scores of 3 to 45%, scores of 4 to 46%, with very few 2s and 5s, and no 1s.

Table 1
A Comparison of Scores Given by Raters

	1 - Low	2 - Below Average	3 - Average	4 - Above Average	5 – High
Rater 1	6 (10%)	23 (38.3%)	16 (26.7%)	15 (25%)	0 (0%)
Rater 2	2 (3.3%)	20 (33.3%)	17 (28.3%)	19 (31.7%)	2 (3.3%)
Rater 3	0 (0%)	0 (0%)	33 (55%)	25 (41.7%)	2 (3.3%)
Rater 4	0 (0%)	3 (5%)	27 (45%)	28 (46.7%)	2 (3.3%)

When raters were asked to provide their rationale for their scores, they reported similar criteria (see Table 2). When describing their criteria for high scores (i.e., 4 or 5), each rater described looking for performances that were unique or expressive. Interestingly, when describing criteria for low scores (i.e., 1 or 2), there were minor disagreements: Rater 1 gave low scores to performances that were "identical or similar to the person next to them," but Raters 2 and 4 based their scores on inability or refusal to follow instructions; Rater 3 did not give any low scores. Regarding average scores (i.e., 3), Rater 1 de-

scribed looking for “small effort to be different... and add something,” but Raters 2, 3 and 4 were more concerned with students taking direction.

Table 2
Responses to Open-Ended Items Regarding Criteria for Creativity

Rater	Criteria for High Scores	Criteria for Low Scores	Criteria for Average Scores
Rater 1	“Something unique (not a wave or clap) and highly unique from the person next to them. I also looked for a level of engagement... was there energy in what they did. It could have been a small movement as long as it was intentional.”	“Was the movement identical or similar to the person next to them?”	“If there was a small effort to be different from the person before and add something that wasn't in the rest of the group.”
Rater 2	“Something that didn't appear to be copied. Willingness to put thought into developing their own movement.”	“Something that didn't appear to be copied. Willingness to put thought into developing their own movement.”	“They were just repeating what several others had done. There was no energy put into coming up with something unique.”
Rater 3	“For creativity, I looked for movement and willingness to step outside of the comfort zone.”	“I hardly gave low ratings, using a small baseline as the average.”	“A nervousness to step out of the comfort zone.”
Rater 4	“I think it was [one particular student] fully committed to the moment he said his name with feeling and his movement was free and personal, it was sharing a piece of himself in a meaningful, playful and lively way.”	“I gave very few 2s and no 1s. [A score of] 2 was based upon commitment and more a calculated refusal to participate.”	“[Scores of 3] were doing the event with the usual nervousness of engaging in this type of work. Students did not plea <i>[sic]</i> a great deal of themselves in the movement and either waved or clapped. The event did not really have much to do with who they were and more about doing what they were asked to do.”

Question 2: Can the CAT distinguish between creativity under different social and environmental conditions?

After determining that judges were qualified to assess creativity, and after establishing IRR, we concluded that we could proceed with a comparison of means. Because of high agreement, we scored each performance based on the averages among all raters and compared the means between the *experimental* group and *control* group. Although neither group had received drama instruction prior to this time, the *experimental* group scored significantly higher than the *control* group. An independent samples t-test below measures the difference between the two groups, and it suggested that there is a possible correlation between environment and creativity of performances, despite both groups having the same teacher and instructions.

The mean score of the *experimental* group was 3.22 (SD = .742) while the mean score of the *classroom* was 2.53 (SD = .789). Additionally, a t-test (see Table 3) revealed that there was a significant effect for setting $t(29) = -3.498$, $p < .001$, with the *experimental* group receiving higher scores. The blackbox was an open space, which allowed for students to perform a greater variety of movements, while the traditional classroom forced students to stand behind their desks, which limited the available space for movement. Many students in the *experimental* group jumped, dived or fell to the floor, but students in the classroom mostly clapped or waved. Additionally, some students in the *control* group attempted big movements, but the instructor prompted them to modify their movements so that their classmates would not collide with their surroundings. Despite both environments allowing students to copy, there was much more copying in the *control* group. Because the creativity difference between the *experimental* and *control* groups was already apparent, the significant difference demonstrated through the t-test provides measurable evidence that the blackbox theatre class had higher creativity scores.

Table 3
A Comparison of Creativity in Black Box vs. Control Environments

	T	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Experimental	-3.498	58	.001	-.692	.198	-1.087
Control	-3.498	57.784	.001	-.692	.198	-1.087

CONCLUSIONS

Regarding the validity of CAT for assessing enacted exercises, it would be hasty to conclude that it is valid for all drama exercises, but we can suggest that it can be used to discriminate between high and low creativity of a warm-up and demonstrate a measurable difference between conditions. The alpha estimate of .819 is strong and suggests that this technique can be useful in the context of drama, but for future inquiry, it may be necessary to consider the effect of verbal instructions on individual performances because, in these recordings, the instructor explicitly allowed students to copy, rather than encouraged them to be original. Consequently, many students copied each other (e.g., “clapped,” “waved”) and this rule most likely affected the way that raters scored performances: three of the four raters (Raters 2, 3 and 4) each reported that they gave scores of 3 to students who took direction, and they scored with the expectation that the baseline for children’s performances would be low. Because they understood that copying was allowed, they gave “average” scores to copied movements, despite their lacking originality - thus, yielding an alpha coefficient that is higher than expected. However, if the instructor had prompted students to try performing movements that their classmates had not performed, then perhaps there would have been more variety in performances, and by extension, there may have been more variance among raters’ scores.

Additionally, we understand that the CAT was originally intended to assess creativity of products created under tightly controlled conditions (Amabile, 1982) and because the “Name and Movement” exercise is open-ended, we will need to evaluate other exercises under different degrees of constraints. However, prior studies have extended the application of the CAT to assess creativity of products created in more naturalistic environments and found that alpha estimates tended to inflate, perhaps because the looser constraints allowed greater variety to be included (Baer, Kaufman, & Gentile, 2004; Kaufman, Cole, Baer, & Sexton, 2008).

Furthermore, we have to consider the meanings behind the scores because it can be difficult to distinguish whether they represent the creativity of the individual (i.e., the student’s unassisted ability) or whether they represent the creativity of the system (i.e., the elevated, or hampered, ability of the student in the setting, social context and instruction). Because this study only gathered one performance at one time point for each student, the scores are not indicative of the individual’s creativity (Oreck, Owen, & Baum, 2003), but instead approximations about social and environmental contexts correlating with creativity. Drama performances include an interdependence of persons, settings and scripts; the scores from this study suggest very little about the creativity of the individual and a great deal about the settings (Barab & Plucker, 2002; Beghetto & Kaufman, 2014) - which are essentially different experimental conditions (Amabile, 1982).

Additionally, if we accept the premise that setting may affect individuals' behaviours in context, then setting may also affect individuals' perceptions as well, including the perceptions of the raters. In drama, the location of art is equally important as the content (Phelan, 2004), so it is possible that performances can appear to be more creative merely because they occurred in a setting more closely associated with creativity. While this will be an important consideration in future studies of the CAT in drama, we are comfortable concluding the effect of setting on raters' scores was miniscule, if any, because raters described the performances in the *experimental* group as having more movement and variety than those in the *control* group.

Because the application of the CAT for this warm-up exercise provides an encouraging start, we were able to use it to compare the creativity of performances from two groups (i.e., *experimental* vs. *classroom*) and we confirmed that the students in the black-box consistently had more creative performances; our application of the CAT supports Amabile's (1982) original intention in creating the instrument, which is comparing creativity under different conditions. This is most likely due to the greater perceived and actual space, which was less constraining than a classroom environment.

STRENGTHS AND LIMITATIONS

Although the CAT was able to measure a difference in creativity between different contexts, the study does not address what features in the environment or social context cause the difference, but instead, it captures the system as a whole. Based on immediate impressions of the blackbox and classroom spaces, we can speculate that the availability of space and, perhaps, pre-existing social scripts influenced children's performances, and by identifying differences in environments, we can isolate features for closer examination in the future. On closer examination, it is also possible that the *experimental* group may have performed more creatively because they entered their first session with the expectation that they would receive future instruction. However, because the mean creativity score of the *experimental* group was one standard deviation higher than the *control* group (*experimental* mean = 3.22, SD = .742 vs. *control* mean = 2.53, SD = .789), we are comfortable in speculating that the space correlates with creativity and we also accept that the difference in student expectations could have had an influence too.

Finally, because the validity of the CAT depends on trustworthiness of the judges, the assessment tool is only as good as the panel (Kaufman, Lee, Baer, & Lee, 2007). Our panel may not have been optimal for assessing all drama performances because we were only able to recruit one "true" expert (Rater 4), but if there was high agreement with the three journeyman actors, then inferences based on their scores should be valid at this

stage. However, this same panel may not be appropriate for assessing creativity in more advanced settings (e.g., theatre conservatories, professional auditions); assessment at advanced stages may require judges to possess more nuanced knowledge, but this panel appears serviceable for assessing creativity in children's drama classes.

With the present findings, we understand that we only examined one exercise and more research on other exercises will be necessary. Although the "Name and Movement" exercise is foundational, it lacks scene work or deep character interpretation. Other exercises should be explored in future work, such as a character monologue or two-person scene. This would impose more constraints on the exercise, which would exclude a number of possibilities. Further research on other drama activities with more controlled and restrictive instructions is needed to better understand the use of this instrument in drama.

Furthermore, because prior studies validated consensual assessment for comparing groups under different conditions (e.g., Amabile, Hennessey, & Grossman, 1986; Baer, Kaufman, & Gentile, 2004; Kaufman, Baer, Cole, & Sexton, 2008), we believe that it remains valid insofar as comparing groups is concerned; however, if we want to compare creativity between individuals, then valid within-subject comparisons would depend on giving each individual several performances in similar environments.

Although we cannot make claims about instruction at this time, we are comfortable suggesting that educators can benefit from using consensual assessment to measure the effect of environment on creativity. Drama educators are vitally interested in preparation, scenery and social interaction, so it would benefit actors and educators to monitor their surrounding conditions to better guide instruction for better creativity. Because the results support the assertion that context correlates with creativity, researchers and actors can conduct follow-up studies to determine the environment that may be optimal. As a starting point, this study suggests that open theatre spaces are better suited for fostering creativity than are constrained classroom spaces. Finally, by making a reliable assessment tool readily available, educators and researchers can coordinate consistent interventions that can monitor and promote favourable environments and social conditions to help students behave more creatively in context.

REFERENCES

- Advares-Yorno, I., Postmes, T., Haslam, S. A., (2006). Social identity and the recognition of creativity in groups. *The British journal of social psychology*, 45(3), 479-497.
- Amabile, T. M. (1982). The social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997-1013.

- Amabile, T. M., Hennessey, B. A., & Grossman, B. S. (1986). Social influences on creativity: The effects of contracted-for reward. *Journal of Personality and Social Psychology*, 50(1), 14.
- Peppler, K., Bender, S., & Catterall, J. (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. Department of Education). Bloomington, IN: Indiana University.
- Baer, J. (2010). Is creativity domain specific?. In the (Eds) J. C. Kaufman & R. J. Sternberg. *The Cambridge Handbook of Creativity*. New York: Cambridge University Press.
- Barab, S. A., & Plucker, J. A. (2002). Smart people or smart contexts? Cognition, ability, and talent development in an age of situated approaches to knowing and learning. *Educational Psychologist*, 37(3), 165-182.
- Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*, 16(1), 113-117.
- Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65-77). IGI Global.
- Beghetto, R. A., & Kaufman, J. C. (2014). Classroom contexts for creativity. *High Ability Studies*, 25(1), 53-69.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, 64(3), 391-418.
- Csikszentmihalyi M. (1988). Society, Culture, and Person: A Systems View of Creativity. In: *The Systems Model of Creativity*. Springer, Dordrecht
- Furman, A. (1998). Teacher and pupil characteristics in the perception of the creativity of classroom climate. *The Journal of Creative Behavior*, 32(4), 258-277.
- Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, 7(2), 193-208.
- Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, 49(3), 234-244
- Isaksen, S. G., Lauer, K. J., Ekvall, G., & Britz, A. (2001). Perceptions of the best and worst climates for creativity: Preliminary validation evidence for the situational outlook questionnaire. *Creativity Research Journal*, 13(2), 171-184.

- Issacharoff, M. (1981). Space and reference in drama. *Poetics Today*, 2(3), 211-224.
- Johnson, C. (2004). Creativity in drama: thinking from within. In R. Fisher & M. Williams (Eds.), *Unlocking creativity: Teaching across the curriculum*. London: David Fulton Publishers.
- Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112-133.
- Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20(2), 171-178.
- Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity*, 2(2), 96-106.
- LeBlanc, A., Jin, Y. C., Obert, M., & Siivola, C. (1997). Effect of audience on music performance anxiety. *Journal of Research in Music Education*, 45(3), 480-496.
- Llewelyn-Jones, L. (2002). Understanding theater space. In *Essays on documenting and researching productions of Greek drama: The sources*.
- Nachmanovitch, S. (1990). *Free play: Improvisation in life and art*. Penguin.
- Neelands, J. (2004). Miracles are happening: Beyond the rhetoric of transformation in the Western traditions of drama education. *Research in Drama Education*, 9(1), 47-56.
- 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.
- Phelan, P. (2004). Marina Abramović: witnessing shadows. *Theatre Journal*, 56, 569-577.
- Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, 39(2), 83-96.
- Podlozny, A. (2000). Strengthening verbal skills through the use of classroom drama: A clear link. *Journal of Aesthetic Education*, 34(3/4), 239-275.
- Renzulli, J. S. (1978). What makes giftedness? Reexamining a definition. *Phi Delta Kappan*, 60(3), 180.
- Sawyer, R. K. (2000). Improvisation and the creative process: Dewey, Collingwood, and the aesthetics of spontaneity. *Journal of Aesthetics and Art Criticism*, 58(2), 149-161.
- Sawyer, R. K., (2003). *Group creativity: Music, Theater, Collaboration*. Routledge Taylor & Francis Group: New York and London.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350.

- Scruggs, M., & Gellman, M. J. (2007). *Process: An improviser's journey*. Northwestern University Press.
- Spicker, H. H. (1992). Identifying and enriching rural gifted children. *Educational Horizons*, 70(2), 60-65.
- Spolin, V. (1999). *Improvisation for the theater: A handbook of teaching and directing techniques*. Northwestern University Press.
- Torrance, E. (1972). Predictive validity of the Torrance Tests of Creative Thinking*. *The Journal of Creative Behavior*, 6(4), 236-262.
- Walker, E., Tabone, C., & Weltsek, G. (2011). When achievement data meet drama and arts integration. *Language Arts*, 88(5), 365.
- Weltsek, G. (2005). Using Process Drama to Deconstruct "A Midsummer Night's Dream". *English Journal*, 95, 75-81.

Corresponding author at: Anthony Phonethibsavads, Santa Ana, CA 92706, USA. E-mail: anphonet@indiana.edu

