Learning Environ Res (2018) 21:43–59
https://doi.org/10.1007/s10984-017-9238-9

ORIGINAL PAPER

# Measures of instruction for creative engagement: Making metacognition, modeling and creative thinking visible

**Christine Pitts**[1] · **Ross Anderson**[1] · **Michele Haney**[1]

Received: 24 June 2016 / Accepted: 22 May 2017 / Published online: 9 June 2017


**Abstract** The purpose of the current study was to estimate reliability, internal consistency and construct validity of the Measure of Instruction for Creative Engagement (MICE) instrument. The MICE uses an iterative process of evidence collection and scoring through teacher observations to determine instructional domain ratings and overall scores. The results demonstrated the sound inter-observer reliability, teacher stability and score validity of the MICE. We found (a) a low proportion of rater variance (0.14–5.99%), (b) moderate to highly correlated within-teacher ratings ranging from $r(17) = 0.663$, $p < 0.01$ to $r(17) = 1.000$, $p < 0.01$ and (c) a statistically-significant difference between classroom teachers and teaching artists, $t(56) = 7.37$, $p = 0.000$. These results relate to the development of classroom environment instruments and the substantive development of pedagogy that supports creative thinking and behaviours, both of which are a priority for enhancing teacher accountability and student learning.

**Keywords** Creativity · Instructional practices · Inter-rater reliability · Teacher evaluation

## Introduction

Generally, classroom observation measures appraise instructional practices to provide formative feedback to teachers on pedagogy aimed primarily at improving students' performance on standardised test scores. Other observation tools target dimensions that evaluate the social and emotional support of classroom environments (Pianta et al. 2010). Few observation tools available to researchers and practitioners identify instructional practices and environmental factors in the classroom that engage students' creatively (e.g.

✉ Christine Pitts
christine_pitts@epiconline.org

[1] Educational Policy Improvement Center, 1700 Millrace, Eugene, OR 97403, USA

Schacter et al. 2006). Yet, research suggests that, by employing creative thinking and behaviours during learning, students adopt real-life problem-solving skills, transfer new knowledge and strategies to authentic situations, build understanding towards improved achievement (Hong et al. 2009) and might construct new meaning (Beghetto 2016). We assert that these creative thinking behaviours are malleable, depend on the learning environment to develop, and could leverage improved academic achievement and social-emotional growth when prioritised in instruction and learning stimuli and modeled explicitly. In this study, we began to fill this gap with one measure that makes visible the fine-grained instructional practices to engage creative thinking and behaviours (Schacter et al. 2006).

## Classroom observation tools

Currently, there is a growing assortment of classroom observation tools applicable for evaluating the school setting and teacher practices, some of which aim to measure similar components of creative thinking and behaviours. For example, the tool and study developed by Schacter et al. (2006) captured the relation between frequency of teaching to creativity and levels of reading achievement. Pianta et al. (2010) *Classroom Assessment Scoring System—Secondary* (*CLASS—S*) is built from a developmental approach in adolescent classrooms by applying an acute focus on teacher–student interactions. Building on the relation of teacher interactions and students' social-emotional growth, our Measure of Instruction for Creative Engagement (MICE) centres learning through creative thinking and behaviours within a learning environment that models openness to different ideas and perspectives. Current learning environment assessments apply to a variety of issues in educational systems, but we found few tools explicitly targeting teacher capacity to develop students' creative self-efficacy, openness and metacognition—potential pre-requisites for students to habituate strategies that foster creative thinking and behaviours in learning (Fraser 1998).

## Proposed design

In designing the MICE tool, we aimed to supplement traditional learning environment measures by using external, objective observations and systematic coding of noted classroom events to evaluate explicit support for creative thinking and behaviours, embedded in typical classroom instruction (Fraser 1998).

### *Creativity and metacognition*

The teaching standards within the MICE tool were synthesised from (a) theory on everyday creativity in learning (Beghetto 2016), (b) metacognitive strategies involved in creative thinking and behaviours (Davis 2000; Hetland et al. 2014) and (c) the potential of modeling creative thinking and behaviours on its development during adolescence (Yi et al. 2015). Aligned with current creativity research trends (Runco 2016), our conception of creative thinking and behaviour employs metacognition for self-assessment and monitoring as a way to measure the opportunities for students to become active learners who are aware of their own skills for making meaning (Davis 2000). In designing the MICE tool, we aimed to measure instruction and modeling of metacognitive practices that intersect with students' opportunities to employ creative thinking and behaviours in their learning (Fraser

1998). In addition, the design of the MICE tool goes beyond the teaching standards from the Schacter et al. (2006) framework, which measured only zero-to-one instructional practice per lesson to support creative thinking and behaviours. For instance, we extended the teaching standards to explicitly describe behaviours for teacher modeling (Ho and Kane 2013) and stages in creative learning (Beghetto 2016).

*Observation framework*

We modeled our tool on the Danielson protocol (Danielson 2013) because it covers a universal framework spanning content areas and school-age settings, dictates much of the learning environment for students, and holds familiarity for practitioners. The instructional subdomains of the Danielson protocol (Danielson 2013) include (a) communicating with students, (b) using questioning and discussion techniques, (c) engaging students in learning and (d) using assessment in instruction which, as an exception to most observation tools, currently have empirical support gathered by multiple teacher evaluation and development systems (Hafen et al. 2014). Similar to our design of the MICE tool, the CLASS—S tool was developed with three broad domains spanning secondary grades and context; yet, that tool focuses more explicitly on teacher–student interactions regarding social and emotional learning and child development, including (a) emotional support, (b) classroom organisation, and (c) instructional support (Hafen et al. 2014; Pianta et al. 2010).

Although the CLASS—S tool (Pianta et al. 2010) and the MICE tool both target metacognition as an indicator of teacher questioning and assessment, the MICE tool explicitly articulates its standards through a lens of creative thinking and behaviours. For example, Hafen et al. (2014) refer to how the CLASS—S tool indicates that teacher analysis and inquiry practices should include higher-level thinking skills commonly used as evidence of student learning—analysis, problem solving and reasoning. In the MICE tool, we aimed for even more specificity by acknowledging creative thinking skills—discussing assumptions, imagining alternatives, taking risks, tolerating ambiguity, openly experimenting and persuading others to understand your view. These distinctions demonstrate the main intention of this study—to develop a tool that captures constructs not traditionally included in existing classroom observation tools.

## Theoretical perspective

Within a systems (Sawyer 2006) and sociocultural (Glăveanu 2013) framing of creativity development in schools, Fig. 1 illustrates the proposed model of creative engagement that guided this study. We referred to Glăveanu's (2013) Five A's—actor, action, artifact, audience and affordances—to classify teachers and students as both actors and audience members engaged in learning transactions. These transactions take place within (a) affordances of the learning environment, (b) actions of the learning process and (c) artifacts produced through learning (e.g. drafts, feedback and final products). This study builds on the assumptions that (a) each of these elements of teacher and student interactions facilitate different dimensions of creativity development as an essential part of the learning process and (b) reliable observation of these interactions furthers conceptual clarity of creative engagement for research and practice (Lench et al. 2015).

In our model the teacher is not only a pivotal *audience* member, but also a crucial creative *actor* whose actions initiate a reciprocal relationship with student creativity—a mechanism to foster the conditions for creative engagement (Glăveanu 2013). According

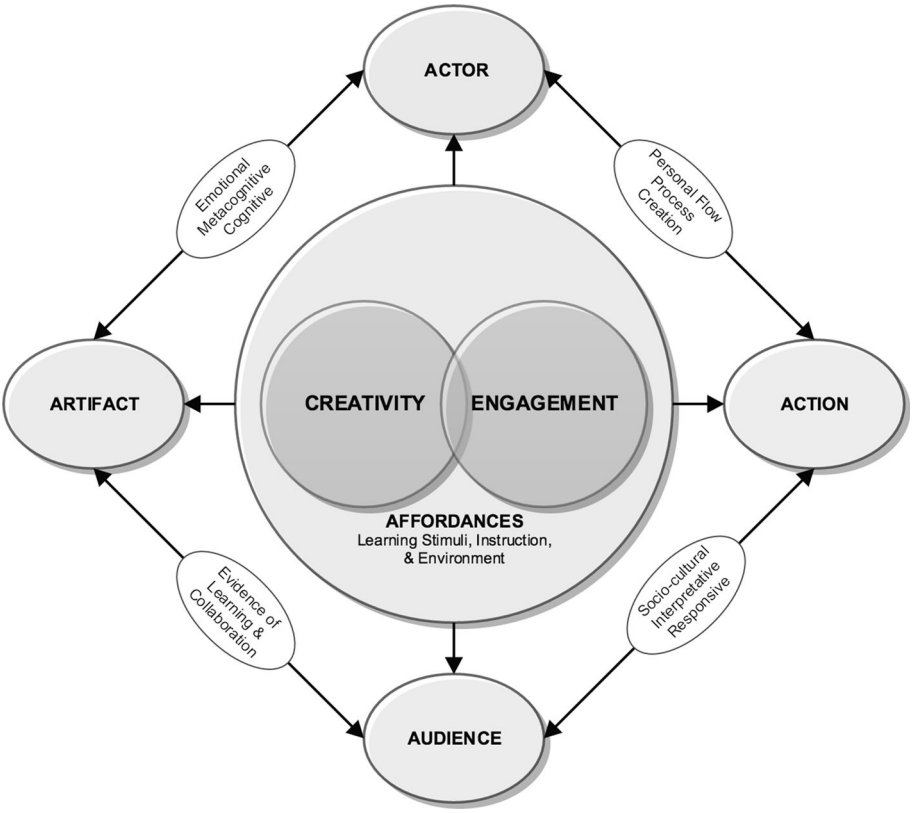


**Fig. 1** Our model of creative engagement in learning that merges Glăveanu's (2013) Five A's conception of creativity as a distinctly socio-cultural process with a synthesis of different elements of student engagement

to Beghetto et al. (2015), students' experience with creativity and learning must be explicitly valued, modeled and instructed in the classroom. Thus, it is imperative that teachers provide experiences that require creative risks, model metacognition and creative thinking and behaviours, and provide opportunities to address ambiguous learning stimuli from multiple perspectives (Glăveanu and Beghetto 2017; Lucas et al. 2013). Furthermore, teachers are responsible for interacting with students during tangible moments of learning (e.g. giving feedback on works-in-progress and improvising mini-lessons). By linking learning environment variables—teacher, classroom and climate—to factors that engage creative thinking and behaviours, our study aimed to explore how domain-general classroom practices enact creativity in the classroom.

The purpose of the current study was to develop the MICE and estimate its reliability, internal consistency and construct validity. To develop and refine the construct of instruction for creative engagement through the MICE, this article addresses the following research questions:

1. To what extent do observers using the MICE observation protocol reliably code and rate the instructional practices for creative engagement?

2. To what extent does the MICE observation protocol provide consistent ratings for the same teachers observed over multiple occasions?
3. To what extent does the MICE observation protocol measure the construct of teaching for creative engagement, during intervention lessons taught by specially-trained teaching artists, compared with core content lessons taught by classroom teachers?

## Methods

This investigation exists within an embedded mixed-methods research design for developing and implementing a school-wide arts integration model. The five intervention schools participating in the larger program of study each collaborated with a professional teaching artist and project leadership team to (a) provide school-wide resources for creative engagement through the arts, (b) design modules with integrated academic and art content and (c) co-teach modules over approximately 6–9 weeks during each trimester of the school year. The measurement study used the MICE observation protocol in all five intervention schools with each of the participating classroom teachers and teaching artists who consented to participate. We conducted observations during teacher-scheduled occasions in the spring and fall of 2015.

### Participants

A convenience sampling method was used to select teachers for observations because only teachers participating in the intervention condition of the study were included in the observations. This study included observations of 21 classroom teachers and 4 teaching artists. The 21 teachers taught in the intervention middle schools participating in the study and the 4 teaching artists all had previous experience teaching art in schools. The teachers taught a range of content areas (i.e. science, social studies, language arts, mathematics and leadership). The teacher and teaching artist sample included 8 males and 17 females. The teachers were in the observations if they were included in the grade level receiving the intervention and consented to participate in data collection. We observed teachers during the spring and fall of 2015, but we do not distinguish between the two time points because we had no theory that predicted that there would be differences in ratings because of the two data collection periods.

### Settings

The intervention took place in five mid-sized middle schools (three city, one suburban and one town fringe) in Oregon that were identified by the state as low performing (Geverdt 2007). The five middle schools, within four school districts, participating in the intervention served high percentages of students identified as minority and low-income, compared with the demographics of Oregon schools and other schools within each district. Of the five schools, the intervention was administered to sixth graders and their classroom teachers in four schools and the seventh and eighth graders attending a single charter school.

Each middle school in the study was identified to be part of the study based on its academic proficiency ratings and more than 50% of their students classified as being eligible for free and reduced-cost lunch. Table 1 describes the demographics of each of the

**Table 1** Descriptive statistics of participating schools from 2013 to 2014 school year data

| Variables | School | | | | |
|---|---|---|---|---|---|
| | School A | School B | School C | School D | School E |
| *N* | 384 | 119 | 334 | 491 | 560 |
| % Students with disabilities | 12 | 14 | 21 | 15 | 18 |
| % FRL | 68 | 62 | 76 | 59 | 84 |
| % English learners | 11 | 0 | 7 | 7 | 21 |
| % African-American | <1 | 0 | 2 | 1 | 2 |
| % American/Indian/Alaska Native | 2 | 2 | 3 | 4 | 2 |
| % Asian | 2 | 1 | 1 | 1 | 1 |
| % Hispanic/Latino | 24 | 9 | 17 | 11 | 29 |
| % Multiracial | 7 | 9 | 8 | 9 | 8 |
| % Native Hawaiian/Pacific Islander | <1 | 0 | 2 | 0 | 1 |
| % White | 65 | 79 | 68 | 74 | 57 |
| % Proficient in mathematics | 57.9 | 40.9 | 56.3 | 45.8 | 53.3 |
| % Proficient in reading | 72.4 | 75 | 75.3 | 63.1 | 64 |

Data retrieved from the Oregon Department of Education School Report Cards, http://www.ode.state.or.us/search/results/?id=116. Percent of student eligible for free/reduced-cost meals is identified by % FRL. Percent proficient in reading and mathematics refers to students who passed the state benchmark on the state assessment in 2013–2014

five schools. All teachers included in the study attended at least one training session as part of the study about arts-integrated teaching for engaging creative thinking and behaviours. The curriculum integrated with art content included science, mathematics, language arts, social studies and physical education.

## Procedures

In the spring of the 2014–2015 academic year and fall of the 2015–2016 academic year, five members of the research team collectively observed 39 lessons for 45–55 min during regular education and intervention lessons. The researchers observed classes that were scheduled by the classroom teacher on regular school days. The content area subjects that we observed included mathematics, science, language arts and social studies, as well as the arts integrated intervention lessons.

## Iterative instrument development, training and qualifications

The purpose of the instrument development and observer training process was to identify the common factors (e.g. the context of the rater, the rater's goals and perceptions, and the rater's understanding of the measurement instrument) that might compromise the instrument's construct validity. During this measurement and evaluation study, we focused on estimating the best inter-rater agreement value necessary for subsequent data analyses.

### *MICE instrument development*

We developed the MICE instrument and standards by converging broad instructional domains adapted from the Danielson Framework and creativity and self-direction

frameworks developed by Lench et al. (2015). Tables 2 and 3 introduce the Question and Engage and Student domains, respectively, and include the attributes and explicit indicators used to score observed practices and interactions. According to its authors, the Danielson Framework (2013, p. 2) was developed to "distinguish between practice at adjacent levels of performance" by articulating instructional practices that encourage students' "deep engagement with important concepts" and aligned pedagogy to the Common Core State Standards. Though the Danielson Framework guided how we framed traditional pedagogical expectations in the MICE, the Essential Skills and Dispositions Framework (Lench et al. 2015) operationalised student and teacher practices for creative thinking and behaviours.

The instructional practices outlined by the Danielson Framework are organised into the five domains of "setting instructional outcomes, designing student assessments, communicating with students, using assessment in instruction and participating in a professional community" (p. 1). When developing the MICE, a team developed then sorted the indicators from the Essential Skills creativity framework within domains and attributes mirroring the Danielson Framework—communicate, question & engage, assess and respond and a final domain specific to actual student thinking and behaviours. We integrated the Skills and Dispositions Framework's five components of creativity, namely, (a) self-awareness, (b) cultivating and evaluating ideas, (c) tolerating risk and ambiguity, (d) experimenting and validating and (e) reflecting and adapting (Lench et al. 2015) into the MICE instrument. Given the developmental level of middle school students, we identified appropriate creativity and self-direction indicators on the beginner-to-emerging expert continuum. Then we translated these indicators into observable practices and interactions that engage students' creative thinking and behaviours.

### *Observer training*

Over 20 h during four collaborative meetings, a team of seven people revised the MICE instrument iteratively to reach a common understanding of intent and content. A team of diverse professionals participated in the development, revision, application and analysis stages of this study. The group included (a) a teaching artist, (b) two former classroom teachers (c) three researchers, (d) a former school administrator and (e) a research scientist with extensive experience in designing and applying observation protocols. During the first observer training in the spring of 2015, the protocol developers shared the tool with the rest of the research team and then made revisions. In the second training session, the research team reviewed the revised version of the protocol looking for clarity and accuracy of the actionable teacher behaviours to provide identifiable evidence of each teaching standard. A smaller subgroup of three researchers met to minimise any redundancies or ambiguity.

For the final training during the spring, each observer used the most-recently revised protocol to score the same two video-recorded lessons independently. The group met to debrief their findings but did not measure their percentage agreement to determine the reliability of the MICE's measurement model. In the final observer training, which occurred in the fall of 2015 during year two of the study, the researchers met to review the protocol used in the spring and to make minor revisions to increase efficiency and clarity.

## Measures

The MICE instrument includes the three phases of (a) literal note taking, (b) identifying indicators of evidence and (c) final scoring on a rubric—each phase is used to determine the

**Table 2** Question and engage domain

| Self awareness and reflection | Cultivating and evaluating ideas | Tolerating risk and ambiguity | Experimenting and validating work |
|---|---|---|---|
| The teacher provides opportunities for, motivates and engages students (e.g. open-ended questions) to… | The teacher provides opportunities for, motivates and engages students (e.g. open-ended questions) to… | The teacher provides opportunities for, motivates and engages students (e.g. open-ended questions) to… | The teacher provides opportunities for, motivates, and engages students (e.g. open-ended questions) to… |
| 1. Explain their thinking to others | 1. Use one's imagination | 1. Engage in unfamiliar experiences, even if they are not immediately successful with them | 1. Explain how their work or the work of others has value |
| 2. Discuss assumptions | 2. Discuss curiosities | 2. Navigate innovation, hesitation and sensible risk taking during their work | 2. Keep an open mind and be imaginative throughout the learning process |
| 3. Reflect on their work and discuss moments of growth | 3. Explore others' ideas and include them in their own work | 3. Share their personal discoveries and challenges with their peers | 3. Publicly self-critique and challenge their own ideas |
| 4. Use their values to make difficult decisions during their work | 4. Explore and discover their own new ideas | 4. Work beyond their personal comfort level | 4. Listen for new insights in others' feedback |
| 5. Explore sources of personal motivation and inspiration | 5. Use imagination to explore alternative possibilities | The teacher…<br>5. Mediates student challenges and models explicit strategies for perseverance | 5. Consider alternative options in their work |
| 6. Accept and discuss their mistakes while they work | 6. Explore content based on personal experiences | 6. Describes and outwardly reflects on their own experiences with taking risks | 6. Respectfully offer constructive criticism |
| | | 7. Provides a safe environment for students to test their answers and solutions | |

domain rating and overall score. First, observers recorded objective literal notes during a classroom observation. Second, the literal notes were used to complete the observation checklist of observable indicators or attributes with the aim of objective documentation and removal of subjectivity. Indicators that were present were marked as evidence of the teaching standards. The rater reviewed evidence and rated on a 4-point scale (1 = 'not or minimally present', 2 = 'somewhat present', 3 = 'present' and 4 = 'developed'). After observers rated each attribute on a 4-point scale, they also subjectively determined the overall domain score on a 4-point scale. The overall domain scores were then summed into global ratings (1–16).

## Analyses

We collected observation data to answer three developmental research questions to validate the reliability and validity of the MICE instrument. First, we established inter-

**Table 3** Student domain

| Self awareness and reflection | Cultivating and evaluating ideas | Tolerating risk and ambiguity | Experimenting and validating ideas |
|---|---|---|---|
| 1. Make their thinking visible and explain their thinking to others | 1. Use their imagination | 1. Explore unfamiliar experiences, even if they are not immediately successful with them | 1. Explain how their work or the work of others has value |
| 2. Discuss their assumptions | 2. Discuss their curiosities | 2. Share personal discoveries and challenges with their peers | 2. Keep an open mind |
| 3. Share and discuss their ideas about learning | 3. Explore and discover new ideas and alternative possibilities | 3. Tolerate less structure in the learning process | 3. Discover new resources, skills and techniques needed to experiment with and communicate ideas |
| 4. Recognize and articulate moments of growth throughout their process | 4. Defer judgement on ideas | 4. Tolerate and learn from mistakes or unintended consequences | 4. Produce prototypes or drafts to understand the reality of ideas |
| 5. Discuss their mistakes | 5. Relate problems or challenges to personal experiences and familiar contexts | 5. Refer to new information and perspectives throughout the learning process | 5. Plan a process for testing ideas and getting feedback |
| 6. Draw on the work of others and own experiences to envision new possibilities | 6. Identify multiple possible directions and consider alternatives with guidance | 6. Contribute to a climate of risk taking and innovation | 6. Commit time and effort to bring work towards completion |
| 7. Seek out and use the feedback of others to think about and plan for their learning process | 7. Eliminate ideas that are not appropriate for context or task | | 7. Remain engaged after failed attempts |
| 8. React to and pursue new opportunities | 8. Refine and elaborate most innovative and impactful choice | | 8. Develop confidence and intention through practice in work |
| 9. Analyse their own work to find meaning and refine and improve ideas and solutions | 9. Gain acceptance of ideas through successful persuasion | | 9. Prioritise choices during creative process based on personal goals and criteria for success |
| | 10. Present ideas independently | | |
| | 11. Collaborate with others to integrate new ideas in their work | | |
| | 12. Incorporate ideas into work that challenge their own ideas | | |

observer reliability of the MICE instrument by measuring the extent to which observers rated the same observable behaviours with the same scores. Assuming that observation scores were reliable, our second goal was to estimate the extent to which the MICE instrument consistently documented observer ratings within observations of the same teacher. Finally, assuming that observation scores were both reliable between raters and within teachers, we aimed to estimate the extent to which the MICE instrument was sensitive to the intervention lessons.

## Reliability

To answer our first research question we analysed inter-observer reliability, which refers to the degree to which individual observers code the same behaviours in similar ways (Smollkowski and Gunn 2012). We calculated the intraclass correlation coefficient ICC (2, k) to estimate the correlation between raters, allowing raters to vary randomly in a two-way random effects design (Shrout and Fleiss 1979).

The variance components of our data were estimated using the procedure within IBM SPSS Statistics, version 21. We calculated the specific ICC (2, k) using the equation

$$\text{ICC } (2,\ 5) = \frac{BMS - EMS}{BMS + (k-1)EMS + k(JMS - EMS)/n'}$$

where BMS is the between mean squared, EMS is the error mean squared, JMS is the judge (rater) mean squared, and $k$ is the number of raters (Shrout and Fleiss 1979). We followed Landis and Koch (1977) recommended guidelines for interpreting ICC reliability estimates: slight reliability (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) and nearly perfect (0.81–1.00). We hypothesised that the ICC would reach a moderate level of 0.41–0.60 given that this study's data were collected as early evidence of the protocol's reliability. We had to achieve power of at least 0.95 to detect an effect size ($\alpha = 0.75$) large enough to reject Type II error with our small sample size ($n = 39$) (Faul et al. 2007).

### *Consistency*

To answer our second research question, we analysed the correlation between the separate observations for each teacher. Following Fan and Sun's (2014) recommendations, we calculated the Pearson correlation coefficient to estimate the reliability of observation scores among teachers.

### *Score validity*

To answer our third research question, we analysed the difference between the average scores of the intervention, ArtCore lessons, compared with the core content lessons. We calculated an independent samples $t$ test and reported its significance and effect size to determine whether the MICE instrument was sensitive to the construct of interest teaching for creative engagement, which we hypothesised would be more present in the intervention lessons.

## Results

Observers from the research team observed and scored 39 lessons taught by 25 teachers and teaching artists. The data were collected during two data-collection time periods in the developmental year of the study, spring 2015, and the first year of implementation, fall 2015. We examined five dependent variables, which included each instructional domain score (i.e. communicate, question and engage, assess and respond, and student behaviour) and the global rating for the MICE instrument. Table 4 presents the descriptive statistics for the dependent variables. We found marginal evidence to support the assumptions of linearity, normality and homoscedasticity. The assumption of normality was marginally tenable because the data included no statistically significant outliers, yet each dependent variable was positively skewed. Linearity also was tenable, although the dependent variables and predictors were slightly correlated. Given a nonsignificant Levene's test for the teacher ratings, $F(56) = 1.537$, $p = 0.220$, the assumption of homogeneity of variances held.

Table 5 presents observations and raters in the data composition matrix. According to Putka et al. (2008), our measurement design was not fully crossed (Rater X Target), nor nested (Rater: Target) making it an Ill Structured Measurement Design (ISMD). In such cases, for the purpose of variance component estimation, a step of determining the ICC, Brennan (2001) recommends viewing the measurement design as fully crossed with increased amounts of missing data. Assuming a fully crossed design, we were missing 60% of our data at random (MAR), because observers were paired with teachers based on available openings in their schedule, not by systematic pairings. Instead of using the traditional ANOVA-based estimation procedures for determining the ICC coefficient, we used restricted maximum likelihood (REML) estimation to reproduce variance components, a practice that is robust to violations of assumptions associated with an ISMD (Marcoulides 1990).

### Descriptive results

Table 4 provides the ranges, means, standard deviations and alpha coefficients estimated for each domain of the MICE and the total global scores. Each domain's scores ranged from the lowest possible score of one to the highest possible score of four. As Table 4 indicates, the four domains, made up of four dimensions, all had good internal consistency (Cohen 1960). The total global scores for the sample ranged from 4 to 14, out of the possible range 4–16, and internal consistency for the total scores was excellent ($\alpha = 0.93$) (Cohen 1960).

**Table 4** Descriptive statistics for the MICE instrument data

| Dependent variable | Range | Mean | SD | Alpha |
|---|---|---|---|---|
| Domain 1: communicate | 1–4 | 2.03 | 1.01 | 0.88 |
| Domain 2: question and engage | 1–4 | 1.81 | 0.85 | 0.86 |
| Domain 3: assess and respond | 1–4 | 1.73 | 0.86 | 0.88 |
| Domain 4: students | 1–4 | 1.73 | 0.80 | 0.90 |
| Global/total rating | 4–14 | 7.30 | 3.19 | 0.93 |

**Table 5** Observation data matrix

| Rater ID | Number of common observations | | | | |
|---|---|---|---|---|---|
| | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5 |
| 1 | 15 | 7 | 5 | 3 | 1 |
| 2 | | 21 | 7 | 7 | 0 |
| 3 | | | 22 | 2 | 8 |
| 4 | | | | 12 | 1 |
| 5 | | | | | 9 |

The values in cells shared by two raters represent their common observations. The values on the diagonal represent the rater's total observations

## Inter-observer reliability

We calculated the intraclass correlation coefficient for our data to test our research question: To what extent do observers using the MICE observation protocol reliably code the instructional practices for creative engagement? Our ICC (2, k) model assumed that differing rater scores for the same lessons depicted rater error and absolute agreement of rater scores, depicting the 'true' score of the teacher's instructional practices observed.

Table 6 provides the variance estimates, proportions and ICC coefficients for each domain and the total global scores of the MICE instrument. Three variance components were estimated, namely, (a) lesson variance corresponds to variation in observation scores between lessons, (b) rater variance corresponds to variation in observation scores between raters, and (c) error variance confounds the interaction of lesson by rater variance and measurement error. The proportion of variance attributable to lessons was high for each domain and the total, ranging from 56.31 to 73.89%. The variance attributable to rater differences was low for each domain and global rating, ranging from 0.14 to 5.99%. The domains that measured teacher communication and student behaviours exhibited the largest proportion of variance attributable to raters, 5.48 and 5.99%, respectively, substantially higher than the other two domains. The variance attributable to the interaction of lesson, rater and measurement error was moderate to high, ranging from 24.50 to 39.81%.

The ICC coefficient alphas estimated were moderate to substantial for each domain and the total (Landis and Koch 1977). Moderate reliability was estimated for Domain 2: Question and engage ($\alpha = 0.60$) and Domain 4: Student behaviours ($\alpha = 0.56$). Substantial reliability was estimated for Domain 1: Communicate ($\alpha = 0.64$), Domain 3: Assess and respond ($\alpha = 0.64$) and the total global scores ($\alpha = 0.73$).

## Teacher stability

We estimated the extent to which observation scores remained stable within teachers across time-points to test our second research question about consistency of ratings within teacher. We calculated the Pearson correlation coefficients between each teacher's individual observations ($n = 2$–6) using the MICE instrument. Our aim was to test the stability of teachers' observation scores and estimate the direction and strength of the relationship among teachers' observation scores. Table 7 reports these correlation data as ranges and averages for each teacher. All of the coefficients were positive and significant at $p < 0.01$, ranging from moderate, $r(17) = 0.663$, $p < 0.01$, to nearly perfect, $r(17) = 1.000$, $p < 0.01$.

**Table 6** Variance component estimates and alpha results for the ICC analysis

| Variance component | Communicate | | Question and engage | | Assess and respond | | Students | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma^2$ | % | $\sigma^2$ | % | $\sigma^2$ | % | $\sigma^2$ | % | $\sigma^2$ | % |
| $\sigma^2_L$ | 0.62 | 64.22 | 0.43 | 59.78 | 0.47 | 64.05 | 0.36 | 56.31 | 7.40 | 73.89 |
| $\sigma^2_R$ | 0.05 | 5.48 | 0.003 | 0.41 | 0.001 | 0.14 | 0.04 | 5.99 | 0.16 | 1.60 |
| $\sigma^2_{LxR,e}$ | 0.29 | 30.30 | 0.29 | 39.81 | 0.27 | 35.81 | 0.24 | 37.69 | 2.45 | 24.50 |
| $\alpha$ | 0.64 | | 0.60 | | 0.64 | | 0.56 | | .73 | |

The $\alpha$ values provided in this table represent the ICC coefficients of the MICE instrument and reflect findings regarding the null hypothesis test

### Score validity

We calculated an independent samples $t$ test to test our third research question about evidence that specialised lessons co-taught by teaching artists demonstrated higher MICE scores then core content teachers' lessons. The arts integration intervention lessons ($n = 12$) were associated with a rating $M = 12.08$ ($SD = 1.78$). In comparison, the regular core content lessons ($N = 46$) were associated with a rating $M = 6.15$ ($SD = 2.62$). Given that the intervention lessons built directly from the fundamental principles of teaching for creative thinking and behaviours, we expected the ratings to be higher for intervention lessons. The independent samples $t$ test was associated with a statistically-significant effect, $t(56) = 7.37$, $p = 0.000$. Thus, the intervention lessons were associated with a statistically-significantly higher mean than the regular core content lessons. At $d = 2.64$, the effect was very large based on Cohen's (1988) guidelines. In other words, evidence from the teaching artists' lessons depicted more teaching practices for creative thinking and behaviours than classroom teachers.

## Discussion

The findings reported in this article include several limitations. First, observations were not fully crossed (Rater × Target), nor nested (Rater:Target), exhibiting an ill-structured measurement design (ISMD) (Putka et al. 2008). Additionally, the analyses were limited by sample size. We were not able to reach the targeted power (0.95) to achieve adequate effect size and reject Type II error in our intraclass correlation calculations. Prior to inferring any conclusions from our results, we suggest further validation studies with a more intentional and stronger research design and larger, more-balanced samples of classroom teachers and artists. Therefore, we caution against overgeneralising from the presented results. Conversely, this study does lay the foundation for future investigation of alternative observation approaches to capture complex constructs at the heart of teaching and learning environments.

### Inter-observer reliability

Congruent with our hypothesis, we established adequate reliability of the MICE instrument with moderate to substantial ICC coefficients across the ratings (Landis and Koch 1977),

**Table 7** Descriptive statistics of Pearson correlation coefficients for teacher ratings

| Teacher | No of observations | *df* | Minimum | Maximum | Mean |
|---|---|---|---|---|---|
| 1 | 4 | 17 | 0.66 | 0.88 | 0.78 |
| 2 | 2 | 17 | 0.85 | 0.85 | 0.85 |
| 3 | 4 | 17 | 0.76 | 0.90 | 0.82 |
| 4 | 2 | 17 | 0.82 | 0.82 | 0.82 |
| 5 | 2 | 17 | 0.84 | 0.84 | 0.84 |
| 6 | 6 | 17 | 0.75 | 0.97 | 0.88 |
| 7 | 4 | 17 | 0.71 | 1.00 | 0.82 |
| 8 | 4 | 17 | 0.90 | 1.00 | 0.93 |
| 9 | 4 | 17 | 0.67 | 0.99 | 0.80 |
| 10 | 2 | 17 | 1.00 | 1.00 | 1.00 |
| 11 | 2 | 17 | 0.79 | 0.79 | 0.79 |
| 12 | 2 | 17 | 0.84 | 0.84 | 0.84 |
| 13 | 4 | 17 | 0.89 | 0.97 | 0.91 |
| 14 | 6 | 17 | 0.77 | 0.95 | 0.87 |
| 15 | 4 | 17 | 0.76 | 1.00 | 0.81 |
| 16 | 4 | 17 | 0.79 | 0.84 | 0.83 |
| 17 | 4 | 17 | 0.78 | 0.97 | 0.88 |
| 18 | 2 | 17 | 0.95 | 0.95 | 0.95 |
| 19 | 2 | 17 | 0.89 | 0.89 | 0.89 |
| 20 | 4 | 17 | 0.86 | 0.96 | 0.90 |
| 21 | 2 | 17 | 0.64 | 0.64 | 0.64 |
| 22 | 2 | 17 | 0.84 | 0.84 | 0.84 |
| 23 | 2 | 17 | 0.94 | 0.94 | 0.94 |
| 24 | 2 | 17 | 0.93 | 0.93 | 0.93 |
| 25 | 2 | 17 | 0.79 | 0.79 | 0.79 |

All correlation coefficients were significant at the 0.01 level

high lesson variance and low rater variance. Our ICC (2, k) model assumed that absolute agreement of rater scores depicted the 'true' score of the teacher's instructional practices observed. Therefore, the high proportion of variance attributable to lessons (56.31–73.89%) provides evidence that the MICE instrument differentiated between different teachers, contexts and content. Congruently, the low rater variance (0.14–5.99%), or the absolute agreement of rater scores, supports our hypothesis that raters using the MICE instrument documented the 'true' teacher practices observed.

The seminal study of teacher observation reliability by Kane and Staiger (2012) revealed higher rater variance in teacher communication (8%) and engaging student domains (6%), which complements the two domains with the largest proportion of rater variance, communication (5.48%) and student behaviours (5.99%). These complementary findings confirm that communication and student behaviour domains are difficult to define and demand conceptual clarity and refinement. The research community invested in classroom observation would benefit from a conceptual and practical model that elucidates observable behaviours attributable to teacher communication and student behaviour.

### Teacher stability

Congruent with our hypothesis that instructional practice for creative thinking and behaviours would remain stable within teachers, we estimated a moderate to nearly perfect correlation that was positive and significant between teachers' scores at multiple time points. The moderate-to-nearly perfect correlations among teacher scores provide some evidence that, if used for high-stakes decisions, MICE observation scores might estimate true instructional practices (Kane and Staiger 2012).

### Score validity

As discussed, ArtCore intervention lessons were associated with a statistically-significantly higher mean than regular core content lessons. The very large effect size $\hat{d} = 2.64$ supports our hypothesis that intervention lessons would register higher scores. Given that they were led by teaching artists and designed to incorporate components that align with the teaching standards of the MICE instrument, this result was not a surprise. The results provide evidence that the MICE instrument measures teaching for creative engagement and that the intervention's potential effect on the learning environment can be differentiated in practice.

### Implications for practice

Our study sought to validate the MICE instrument by connecting teacher, classroom and climate variables with student creative thinking and behaviours in learning. Kane and Staiger (2012) recommend broadening the focus of classroom observation tool research from the narrow view of inter-observer reliability towards determining the best practices regarding (a) number of lessons, (b) number of observers and (c) different students. Our proposed instrument broadens the substantive and technical scope by measuring the pedagogical practices that afford students a learning experience focused on activating creative learning. We suggest that the MICE instrument provides information about instructional choices not captured by traditional teacher observation measures and could prove useful for measuring the quality and quantity of teacher practices to creatively engage students and for providing meaningful feedback. As states begin redesigning locally-developed accountability models that align with the federal Every Student Succeeds Act (ESSA), it is incumbent to use measures that efficiently measure instructional practices related to students' deeper engagement and creativity in learning (Darling-Hammond et al. 2016). Our study provides an initial indication that we can reliably and validly measure such instructional practices.

### Implications for research

Overall, the results demonstrate promise of inter-observer reliability, teacher stability and score validity for the MICE instrument. Based on the promise of our findings and their relation to practical application in evaluating and refining innovative pedagogical models, use of the MICE instrument in a variety of contexts (e.g. elementary and high schools) will be needed to consider further generalisability. Other researchers using observation tools to study creativity in the classroom might consider (a) the appropriate measurement design (e.g. fully nested or fully crossed), (b) an increased sample size, (c) running a *G Study* analysis to disaggregate the residual facet and (d) new ways to describe complex constructs

through observable facets. We found that the variance attributable to the interaction of lesson and rater and measurement error was moderate to high, ranging from 24.50 to 39.81%. This confounded variance component of the interaction and error was similar to what Kane and Staiger (2012) found when estimating variance to determine the reliability of similar instructional domains. In the future, we recommend following the advice of Ho and Kane (2013) of separating the lesson by rater interaction from the residual to determine substantive differences of the interaction between raters and lessons or teachers. As researchers begin to hone in on non-academic factors that could predict a host of critically important outcomes (e.g. dropout), measuring the classroom conditions that support student creativity, metacognition and agency in meaningful learning will be critical. Our study provides one more step in that direction.

## Conclusion

Our results demonstrate that the MICE instrument provides information about teaching practices for creative thinking and behaviours that traditional observation tools generally fail to capture. In addition to measuring instructional choices reliably, the MICE tool appears to evaluate the frequency and quality of opportunities for students to engage their creativity and metacognition in their learning. Future work on the MICE protocol will focus on closing the feedback loop to affect teachers' own metacognition and creative potential. We estimated that the MICE instrument observation data were moderately to substantially reliable for determining true instructional differences, with little effect from raters. Additionally, we found that the MICE instrument documented stable scores within teachers and therefore, if the measure were used on multiple occasions, it probably would provide similar scores for the same teacher without intervention. Finally, the results provide evidence of construct validity to capture targeted practices. As such, our results indicate that the MICE can support that critical component in the cycle of improvement to enhance the conditions for creativity in the classroom—effective measurement.

**Acknowledgements** This research was supported by a grant from the U.S. Department of Education (PR/Award No. U351D140063).

## References

Beghetto, R. (2016). Creative learning: A fresh look. *Journal of Cognitive Education and Psychology, 15*(1), 6–23.

Beghetto, R., Kaufman, J., & Baer, J. (2015). *Teaching for creativity in the common core classroom*. New York: Teacher's College Press.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Cohen, J. (1960). A coefficient for agreement of nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. Princeton, NJ: The Danielson Group.

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., et al. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. Palo Alto, CA: Learning Policy Institute.

Davis, J. H. (2000). Metacognition and multiplicity: The arts as models and agents. *Educational Psychology Review, 12*(3), 339–359.

Fan, X., & Sun, S. (2014). Generalizability theory as a unifying framework of measurement reliability in adolescent research. *Journal of Early Adolescence, 34*(1), 38–65.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Fraser, B. (1998). Classroom environment instruments: Development, validity and applications. *Learning Environments Research, 1,* 7–33.

Geverdt, D. (2007). *Remote towns and rural fringes: An overview of the NCES School Locale Framework*. Washington, DC: U.S. Census Bureau. http://aasa.org/uploadedFiles/Policy_and_Advocacy/files/RemoteTownsRuralFringes.pdf.

Glăveanu, V. P. (2013). Rewriting the language of creativity: The Five A's framework. *Review of General Psychology, 17*(1), 69–81.

Glăveanu, V., & Beghetto, R. (2017). The difference that makes a 'creative' difference in education. In R. Beghetto & B. Sriraman (Eds.), *Creative contradictions in education* (pp. 37–54). Cham: Springer.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2014). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the classroom assessment scoring system-secondary. *Journal of Early Adolescence, 35*(6), 650–680.

Hetland, L., Winner, E., Veenema, S., & Sheridan, K. (2014). *Studio thinking 2: The real benefits of visual arts education*. New York: Teachers College Press.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (Research Paper, MET Project). Seattle, WA: Bill & Melinda Gates Foundation.

Hong, E., Hartzell, S. A., & Greene, M. T. (2009). Fostering creativity in the classroom: Effects of teachers' epistemological beliefs, motivation, and goal orientation. *The Journal of Creative Behavior, 43*(3), 192–208.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Research Paper, MET Project). Seattle, WA: Bill & Melinda Gates Foundation.

Landis, J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. doi:10.2307/2529310.

Lench, S., Fukuda, E., & Anderson, R. (2015). *Essential skills and dispositions: Developmental frameworks for collaboration, creativity, communication, and self-direction*. Lexington, KY: Center for Innovation in Education at the University of Kentucky.

Lucas, B., Claxton, G., & Spencer, E. (2013). *Progression in student creativity in school: First steps towards new forms of formative assessments* (OECD Education working Papers, No. 86). Paris: Organization for Economic Cooperation and Development/OECD Publishing.

Marcoulides, G. A. (1990). An alerternative method for estimating variance components in generalizability theory. *Psychological Reports, 66,* 102–109.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2010). *Classroom Assessment Scoring System—Secondary (CLASS—S)*. Charlottesville, VA: University of Virginia.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93,* 959.

Runco, M. A. (2016). Commentary: Overview of developmental perspectives on creativity and the realization of potential. *New Directions for Child and Adolescent Development, 151,* 97–109. doi:10.1002/cad.20145.

Sawyer, R. K. (2006). Explaining creativity: The science of human innovation. New York: Oxford University Press.

Schacter, J., Thum, Y. M., & Zifkin, D. (2006). How much does creative teaching enhance elemtary school students' achievement? *Journal of Creative Behavior, 40,* 47–72.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Smollkowski, K., & Gunn, B. (2012). Reliability and validity of the Classrooom Observation of Student–Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly, 27,* 316–328.

Yi, X., Plucker, J. A., & Guo, J. (2015). Modeling influences on divergent thinking and artistic creativity. *Thinking Skills and Creativity, 16,* 62–68.