Reliability of Pre-Service Physical Education Teachers' Coding of Teaching Videos Using Studiocode Analysis Software

Keven Prusak
keven_prusak@byu.edu

Brigham Dye

Charles R. Graham
charles.graham@byu.edu

Susan Graser

Follow this and additional works at: https://scholarsarchive.byu.edu/facpub

Part of the Educational Psychology Commons

BYU ScholarsArchive Citation
Prusak, Keven; Dye, Brigham; Graham, Charles R.; and Graser, Susan, "Reliability of Pre-Service Physical Education Teachers' Coding of Teaching Videos Using Studiocode Analysis Software" (2010). Faculty Publications. 840.
https://scholarsarchive.byu.edu/facpub/840

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Reliability of Pre-Service Physical Education Teachers’ Coding of Teaching Videos Using Studiocode® Analysis Software

KEVEN PRUSAK, BRIGHAM DYE, CHARLES GRAHAM, AND SUSAN GRASER

Brigham Young University
keven_prusak@byu.edu
brigham.dye@gmail.com
charles_graham@byu.edu
sue_graser@byu.edu

This study examines the coding reliability and accuracy of pre-service teachers in a teaching methods class using digital video (DV)-based teaching episodes and Studiocode analysis software. Student self-analysis of DV footage may offer a high tech solution to common shortfalls of traditional systematic observation and reflection practices by increasing the amount, timeliness, and accuracy of performance feedback. What is yet to be determined is whether students can reliably and accurately analyze such footage. Using Studiocode software, student analyses were compared to those of experts to determine coding reliability and content accuracy. The results of this study indicate that with less than two hours of training and three practice attempts, students are moderately reliable in their coding ability and highly accurate in their content analysis. Students who engage in additional attempts demonstrated high levels of coding reliability and content accuracy. Implications of this study include (a) students can
reliably learn to self-code within a reasonably short period of time—making these technologies manageable in teaching methods courses—and (b) DV analysis may provide additional, accurate, and reliable sources of feedback beyond traditional evaluative techniques.

Key Words: physical education teacher education, video analysis, reliability, student teaching, feedback, self-reflection

Because teaching is a performance-based profession, teacher education programs commonly integrate classroom-based instruction with field-based practice. Typically, this integrative process is one in which pre-service teachers systematically and increasingly assume responsibility for more complex teaching events. For example, pre-service teachers may begin their field experiences engaged in peripheral observation of classroom interactions. Then, in partnership with an experienced cooperating teacher, they would begin presenting discrete parts of a lesson. Eventually, the teacher candidate would assume full responsibility for the management and instruction of the classroom while the cooperating teacher assumes a more peripheral role as observer/evaluator.

The purpose of this integrative, transitional process is to provide timely support to pre-service teachers as they practice and develop increasingly complex teacher competencies and characteristics. Accurate and timely feedback throughout this process facilitates pre-service teachers’ acquisition of these desired competencies.

Feedback in field-based experiences, typically resulting from peripheral observational techniques such as systematic observation, field notes, or checklists, comes from a supervising teacher-educator or the cooperating teacher (Darst, 1989). While these sources of feedback can help focus pre-service teachers’ practices, they are often fraught with several limiting realities such as frequency, cost, recall accuracy, memory decay, and disparity in student- vs. observer-perception vs. reality (Sharpe, 1997; Sherin, 2004). Alone, observational techniques likely fall short in their intention to maximize the quality and quantity of timely feedback to pre-service teachers during their development period (Struyk & McCoy, 1993). To augment observational techniques, teacher educators have engaged pre-service teachers in self-reflection exercises following teaching episodes. Becoming a capable reflective teacher has been a valuable additional source of feedback for pre-
service teachers but still suffers from many of the same obstacles associated with a peripheral observation.

The Challenge of Providing Accurate and Timely Feedback: Limitations of Current Practices

Coursework in teacher education programs often focuses on the development of pedagogical content knowledge (PCK, the blending of pedagogical and content knowledge). It involves how specific topics, problems, or questions can be organized, represented, adapted, and taught to learners (Shulman, 1987). Classroom learning of pedagogical content knowledge may be facilitated by a number of learning activities ranging from textbook readings to video-based case studies. In general, pre-service teachers’ pool of pedagogical content knowledge developed through non-performance-based learning remains largely theoretical. Disconnected from performance-based knowledge, this pool of PCK tends to be somewhat inert. Although pre-service teachers may be able to explain this inert knowledge when explicitly prompted to do so, these teachers do not spontaneously activate this relevant knowledge when faced with a practical situation in which it is potentially useful. Engaging students in early and frequent performance-based practice, such as in field experiences, is thought to help transform this inert, declarative knowledge into viable, performance knowledge.

For most pre-service teachers there exists a misalignment between PCK and actual teaching performance. When first engaging in field-based experiences, a large portion of actual teaching performance initially aligns with personal assumptions and prior education experiences then with the body of PCK developed in classroom-based coursework (Winograd, Higgins, McEwan, & Haddon, 1995). Appropriate feedback can serve to highlight and resolve the misalignment of theory and practice for pre-service teachers and thereby facilitate their move toward becoming master teachers. As theory and practice become ever more aligned, a convergent dialogue occurs wherein theory informs and shapes practice in an attempt to solidify theoretical concepts. Through iterative cycles of practice followed by accurate feedback, reflection, and planning, pre-service teachers should experience increased alignment and ongoing convergent dialogue (between declarative and performance knowledge) hopefully resulting in improved teaching performance.

It would be reasonable to assume that an increase in the quality and quantity of timely feedback would lessen the number of needed iterations
and would, in effect, “shorten the learning curve.” From a constructivist perspective (Karppinen, 2005; Rovegno & Dolly, 2006), new information is processed based upon one’s experiential background, which is strongly influenced by personal perceptions of past and new experiences. Once processed, new experiences become a part of one’s experiential background. Thus, simultaneously, as one’s pool of PCK increases, or perhaps is dramatically reorganized, practice attempts will begin to more fully align with that PCK. Therefore, improved perceptions resulting from accurate and timely feedback are of paramount importance.

In practice, the impact of feedback in teacher education is hampered by perceptual and communicative breakdowns. Often, external evaluators and practicing teachers have their own unique perspectives about the lesson taught. This is not surprising considering the depth and breadth of their respective experiential backgrounds. Through experience, expert teachers are able to focus on the needs of individual learners, automate classroom routines, develop flexible long-term plans, transition smoothly from one activity to another, integrate assessment into instruction, focus on student comprehension, organize instructional content clearly, and attribute successes/failures to controllable factors (O’Donnell, Reeve, & Smith, 2007).

Further, perceptions dull very quickly following a lesson (i.e., memory decay), particularly on the part of the pre-service teacher who, of necessity, must dedicate so much attention to the execution of the lesson. For example, pre-service teachers are often so completely focused on the mechanics of the lesson delivery and classroom management that they are often unaware of its effect on student learning and behavior (Heibert, Gallimore, & Stigler, 2002).

However, while these student effects, as well as the pre-service teacher’s performance, are often evident to an external evaluator (Berliner, 1992; Bransford, et al., 2006; Hogan, Rabinowitz, & Craven, 2003), they can hold little meaning when later discussed by the observer and practicing teacher (i.e., disparate perceptions). Consequently, when pre-service teacher and evaluator attempt to debrief, their ability to communicate is limited by the degree to which they share a common perception of the teaching performance. Feedback messages are complex in nature, requiring active interpretation (Nicol & Macfarlane-Dick, 2006) on the part of the student and, again, are limited by the amount of shared perception. In those areas where shared perceptions are high, the feedback becomes more meaningful. With disparate perceptions little or no meaning is conveyed and frustration for both parties can result.
The external evaluator is also hampered by the necessity to selectively focus on certain aspects of the lesson while ignoring others. Systematic observation techniques, for example, attempt to quantify certain aspects through the use of event coding, duration, frequency, etc., none of which tell the whole story. Furthermore, external evaluators are often unaware of contextual history and needs specific to a particular setting and how these may affect the lesson being observed. In other words, it is impossible for any external evaluator to capture, synthesize, and debrief the entirety of any lesson (i.e., recall quantity). This is not to say that such evaluations lack value; rather, it highlights their limitations.

Memory decay, differing perceptions, differing experiential backgrounds, incomplete synthesis, and lack of meaningful communication all limit the accuracy and value of external feedback or reflection exercises. In short, the common practice of an external evaluator sharing feedback with a pre-service teacher while debriefing after a lesson is hampered by an incomplete comprehension on the part of both parties. Increasing the perceptual accuracy of both teacher and evaluator, as well as their ability to communicate about those perceptions, would be valuable in teacher education.

The Digital Video Solution

Digital video analysis may provide a number of instructional affordances not otherwise available in teacher development (Brophy, 2004; Chan & Harris, 2005). It may also provide a way to reduce perceptual and communicative breakdowns in the reflection (Tripp, in review; Wright, 2008) and evaluative feedback process. First, while using video footage in teacher education is not new, it has rarely been used to its potential because of the cumbersome nature of previous analog technologies. The advantages of DV technology include the ability to instantaneously manipulate captured images, thus facilitating the timely extraction of focused, quality feedback. Also, DV images are easily distributed as DVD’s or web-streaming. Second, DV software tools remove the temporality of the performance. That is, captured DV images are not subject to memory decay, perceptual differences, inaccurate recall, or communicative breakdowns. They are high fidelity evidence in an unaltered form. They can be viewed and reviewed through a variety of user-defined perceptual lenses (such as, evaluator-, teacher- or student-perspective) and reasons (such as improvement of teaching performance or assessing student learning). Third, DV analysis tools provide a shared representation of the performance, anchored in specific events, around which
performer and evaluator can communicate. Rather than speak in generalities about teaching performance, both can direct one another’s attention to particular events that represent their respective points of view. Fourth, DV analysis may provide a cost-effective means of increasing the quality and quantity of feedback for pre-service teachers during their developmental period. It is unrealistic to expect that one college supervisor can observe and provide feedback to each and every practicing teacher. Providing evaluative feedback, therefore, tends to be a haphazard approach. For example, classroom circumstances during the evaluator’s visit may not occasion a specific teaching-learning interaction for which feedback is needed. Instead of having an evaluator observe and debrief 30 students, DV could allow 30 students to analyze themselves every time they taught. These individual analyses can receive further analysis and feedback from the instructor as needed. While providing quality feedback is a desirable goal when an evaluator arrives to observe, these visits often result in heightened anxiety. Digital video can provide a dispassionate, impartial view of the lesson that relieves much of this anxiety.

Digital video analysis tools suggest the possibility that performers can generate effective self-evaluative feedback without necessitating an external evaluator. The purpose of an external evaluator is generally to provide a more objective perspective on the performance. This means the external evaluator does two things. First, since the external evaluator is not engaged in teaching, he/she has more processing capacity to notice things that might escape the attention of the performer. Second, the external evaluator often brings a more complete understanding of the relevant pedagogical content knowledge, which helps the expert evaluator notice things that the pre-service teacher could not notice. For pre-service teachers to provide sufficient self-evaluative feedback, they would have to be able to perform these two functions of the external evaluator.

Digital video analysis tools make it possible for the performer to observe the performance after the fact. This means performers can notice aspects of their performance to which they were blind while in the act of performing. Thus, the performer can fulfill the first function of an external evaluator. In order to fulfill the second function of an expert evaluator, pre-service teachers would have to be trained to notice aspects of the teaching performance that would be noticed by an expert evaluator.

Reviewing video footage as a source of feedback is not without its peculiar drawbacks (i.e., cumbersome VHS technologies). Further, novice teachers, according to Brophy (2004), do not seem to get much benefit from self-guided video inquiry alone (i.e., simply viewing video is not enough to
Accuracy of Pre-Service Teachers Video Self-Analysis

Develop expertise). Perhaps combining self-guided video inquiry with attentional cuing (learning and applying a set of codes representing a pre-defined list of desired competencies) could provide essential scaffolding by directing student attention to relevant and important aspects of the video. Brawdy and Byra (1994), demonstrated that within a set of well defined criteria, pre-service teachers themselves can develop the skills necessary to become a valuable source of personal feedback. Essentially, a new skill set, necessary for accurate “noticing” "(i.e., viewing with a trained eye) (Bransford, et al., 2006; Sherin & vans Es, 2005)" is created.

If video analysis helps pre-service teachers provide themselves with reliable and valid self-evaluation, the costs of and challenges of relying solely on external observation and reflection practices (Cunningham, 2002) may be ameliorated. In other words, DV analysis may increase recall accuracy and quantity, align students and observer perceptions, provide realism as well as increase cost-effectiveness for teacher education programs. If realized, these benefits of DV as a feedback source may allow pre-service teachers to engage in more focused and effective practice during their developmental period.

However, before researchers can investigate the effects of DV analysis on performance, they need to establish whether students can analyze their own videos in a valid and reliable manner. The extent to which pre-service teachers in a Physical Education Teacher Education (PETE) course can be taught to reliably apply a set of expert-defined codes to videos of their own teaching performance using DV analysis software remains unclear. Therefore, the purpose of this study is to examine the degree to which pre-service teachers can reliably code teaching episodes when compared to experts.

Preliminary work in physical education teacher education (Prusak & Wilkinson, 2007), provide three strategies for the use of DV analysis in training new teachers namely, (a) video-aided systematic observation techniques, (b) video-aided reflection exercises, and (c) video-based performance analysis. Employing these strategies, PETE instructors at a large western university in a physical education course for elementary school teachers used a video analysis tool and a set of clearly defined codes associated with recognized best teaching practices to provide a scaffold to this noticing process. Prior to coding their own videos, however, pre-service teachers are trained in the meaning and application of these codes while applying the codes to videos of expert PE teachers. Thusly trained, pre-service teachers are then asked to code their own videos.
What remains to be determined is the degree to which pre-service teachers can arrive at similar evaluative conclusions as would their experienced evaluators in the analysis of their own videos. Therefore the purpose of this study is to assess the coding reliability of pre-service teachers as compared to the analysis of experts as they (a) code an expert teacher and (b) code their own teaching performances.

Only if we first know that students can be trained to reliably analyze their own teaching can we be assured that such DV analysis practices can increase the quality and quantity of timely feedback over and above observational and reflective practices. If students can be shown to be reliable self-analyzers, then the question of whether this type of feedback leads to improved performance can be examined. The findings of this study could potentially lead to improved feedback practices resulting in more deliberative practice and ultimately to better teaching.

METHOD

Participants

Participants (n = 49, one male and 48 females) for this study were recruited from two concurrent sections (Section A, n = 23 and Section B, n = 26) of a Physical Education for Elementary School Teachers course at a large western university. All participants signed informed consent forms agreeing to participate. The university Institutional Review Board approved all procedures.

Instructional Context And Intervention

The two professors responsible for the development and implementation of various DV analysis software protocols in their courses each taught one of the two sections. This two-hour course is taught on a block schedule (twice weekly for two hours over eight weeks). This schedule allows for the integration of four weekly field-based teaching experiences in which students team-teach three times and a fourth time alone. This course is most often taken in the first semester as students enter the elementary education program and therefore is likely to be their first exposure to teaching in an elementary school setting. This format presents challenges to the instructors
who must cover a considerable amount of curriculum heavily focused on management and instruction in PE in the first five class periods in order to prepare the students.

An integral part of the course involves training students to use the DV analysis software tool, Studiocode (see Figure 1), to analyze video of their own teaching episodes. Studiocode integrates DV editing capabilities with a user-defined coding system allowing user to code (i.e., attach meaning to selected instances) to video footage. Once coded, instances can be manipulated in a variety of ways, including instantaneous recall as well as adding written reflections via a transcription tool.

The instructors created a list of desired teaching competencies and associated codes specific to the objectives of this class to be used when analyzing videotaped lessons. Students are given this list on the first day and instructed to refer to it often as they plan their lessons so that they can demonstrate all of the competencies during their field experiences. This list of competencies while not exhaustive was created after careful study and consideration of the best practices of master elementary PE teachers to suit the constraints of this course. These competencies focus on essential management and instructional strategies specific to the activity-based elementary PE setting. They are intended to provide students with enough skills to perform at the level of a student teacher with the realization that continued improvement is needed to reach teaching mastery.

A set of five codes and labels (descriptors) correspond to five specific desired competencies (see Appendix) as follows: (a) freeze position (either compliant or non-compliant); (b) equipment transitions (moving equipment on to or off of the floor); (c) student transitions (toe-to-toe, splitting class, and whistle mixer); (d) instruction (“when-before-what”, short- and long- instructions); and (e) discipline plan (six steps in the plan). In all, students would code and label fifteen discrete behaviors (five codes plus their labels) as they occur in a videotaped lesson. Since it is unlikely that any one lesson would contain examples of all 15 instances, students would tape each of four lessons they taught to assure complete coverage of the list of desired competencies.

Students first practiced using Studiocode to analyze videos of an expert teacher followed by applying those skills to the coding of the videos from each of their four teaching episodes.

Studiocode condenses the coding information into a matrix that summarizes the number of codes and their associated labels. Following each coded lesson, students submitted a printed copy of the matrix as evidence of a timely completion of the analysis. Once all four lessons are coded, the stu-
Students sorted through all instances and compiled their best examples of each of the desired competencies to be submitted electronically to a database for instructor review and grading. Using a pre-established rubric, actual teaching performances and demonstrated degree of competency of the skills in the desired competencies are evaluated.

Data Collection and Analysis

This study included two phases. First, students participated in a coding proficiency test (CPT) using Studiocode to analyze a video of an expert teacher leading a PE lesson to determine how reliably students could code. Second, researchers conducted a content audit of the students’ analysis of their own teaching videos to determine the accuracy of applying codes and label correctly.

Coding proficiency test. During the first phase of data collection, researchers conducted a CPT that compared the coding abilities of student teachers to those of expert PE teacher educators to assess the reliability with which students could code a lesson. A video of an expert teacher delivering a PE lesson that included the demonstration of a majority (13 of a possible 15 desired competencies were evident) was selected then independently coded by the instructors. Results from the independent coding of the video were compared via the matrices. This allowed the instructors to compare the number of codes and labels each used while coding the video. The instructors then discussed discrepancies between their coding matrices and resolved ambiguities in how they had applied the coding scheme. Based on this discussion, they agreed upon a single Studiocode file that would serve as the standard for expert coding for the video.

During the first few weeks of class, students received intensive training on the teaching competencies represented in the competencies list and coding scheme. Students both viewed several videos of expert teachers as well as participated in the live modeling of the desired competencies. Live modeling sessions were held on the gym floor where the instructor and students role-played as if it were an actual PE lesson complete with equipment, music, and instructional signs.

Though the two professors had worked together for several years prior to this study, additional collaborative planning sessions were held to assure uniform instruction about the desired competencies. Minor discrepancies in the competencies were resolved before presenting them to the students in class lectures and in a handout. Ongoing dialogue helped to ensure both classes were receiving consistent information.
Additionally, the primary researcher attended both classes to observe how each professor presented the information about the desired teaching competencies. Field notes recorded any differences in course content delivery that may have influenced how the students may have understood the competencies, Studiocode or the proficiency test. Constant, daily comparisons between field notes from each setting were compared and differences noted in content covered on which days. Listing significant similarities and differences between the two classes further augmented analysis.

In addition to instruction specific to the desired competencies, students were trained in the use of Studiocode software including basic keystrokes, capturing, coding and databasing video footage. Following this initial training, the students participated in a coding proficiency test. Students were asked to code the previously selected and coded video of the expert teacher. The initial coding test could be done in approximately one hour and subsequent attempts varied from a few minutes up to 30 minutes. They were informed that the intent of the proficiency test was to compare their coding to that of the experts and that points would be deducted if they miscoded or mislabeled any event. A goal of 80% agreement was set and students were allowed to revisit their coding as many as three times if they fell short. They were informed that their grade was not tied to how well they coded the video but rather for completing the test. Apparently for students in Section A, who were the first to complete the test, this was unclear and it was noted that students were retaking the test as many as six times in order to improve their scores. While this confusion caused some minor frustration, it serendipitously yielded some important findings pertinent to this study. Section B instructor gave particular emphasis to students to not take the test more than three times.

Scoring of the proficiency test was accomplished by exporting the Studiocode data (time-line signatures) into a spreadsheet programmed to compare the students’ codes with those of the experts. This procedure yielded two percentages, one describing the percentage of agreement and the other describing the percentage of overlapped time of the coded instances between student and expert. The first score was an indication of correctly coding/labeling an instance and the second an indication of students ability to recognize the competency in its entirety. Additionally, CPT results were immediately provided for each of the 15 competencies as to whether they had too many or too few of each code. With this specific feedback as to where their errors occurred, students could revise their analysis with greater precision and resubmit subsequent analyses. All attempts were labeled attempt 1, attempt 2, and so on and were saved for analysis and comparison.
Descriptive statistics, broken down by class, were examined for the proficiency test. In a few cases, some students did not complete a minimum of three attempts and thus the different $n$'s (see Table 1).

As expected, the mean scores for each successive attempt showed improvement and these differences were examined using within-group repeated measures ANOVA but only for those students who completed all three attempts. Between group (class) differences were assessed using a t-test.

Descriptive statistics on each item were analyzed to identify any specific codes that may have been problematic for students. Mean coding errors for each item were calculated and those with a 40% error rate (or the highest tertile) were deemed problematic.

After both classes had completed the proficiency test, follow-up interviews with students and instructors were conducted. Students were divided into tertiles according to test scores and four students from the highest tertile and four from the lowest tertile were selected for follow-up interviews. Three were from class A and five were from class B. This purposeful sampling assured a range of student experiences with the proficiency test. During the interviews instructors were asked to explain why students might have coded the video differently than the experts. Interview notes were analyzed using simple qualitative methods to explore reasons for the lower-than-anticipated scores on the proficiency test.

**Content audit.** The second phase of data collections involved gathering information about how well students correctly identified the competencies as they occurred in their own teaching. Videos used for this self-analysis phase were those captured during the solo-teaching day. A fellow student filmed the event and was instructed to keep the teacher in the video frame while also capturing as many of the students in the frame as possible. Each student wore a remote microphone to assure clear audio.

Students then used StudioCode software to code their own teaching videos using the same list of 15 desired codes and labels as in the proficiency test. To determine the content accuracy of coded instances (i.e., coded instances clearly represented the intent of the codes and labels used), a random sample of 20 students (10 from each section) were selected for further analysis. For each of these students, 15 randomly selected instances were examined as to whether they were accurate depictions of codes and labels used by the student. The researcher coded each instance as either “correct” or “incorrect.” If coded as incorrect, an explanation as to why was added.
RESULTS

Reliability of Expert Coding

To establish the standard by which the participants would be assessed, the two expert coders each separately coded a lesson taught by a master teacher using the list of desired competencies previously agreed upon. Initial comparisons revealed that the two experts agreed 76% of the time. Upon examination, the discrepancy was found to lie in the interpretations of one of the codes and its label (giving short instructions) and one missed instance of using the discipline plan (second time out). Once these discrepancies were discussed and alleviated, a score of 96% agreement with 92% overlapping time was achieved. Out of the 59 possible coded instances, only four showed any discrepancies between expert coders. These scores demonstrated a high degree of interrater reliability as well as assuring that subsequent explanations regarding the list of desired competencies would be consistent.

Coding Proficiency Test Results

High interrater reliability is traditionally but somewhat arbitrarily set at 80%-85% and initial hypothesis supposed that this was attainable by the students in the present study. In this study, class means reveal reliability scores ranging from a low of 43% for Section B on attempt 1 to a high 76% for Section A on attempt 3. Overall means following the third attempt indicated moderate levels of reliability for matching codes (63%) and overlapping time (68%; see Table 1 for full results). While falling short of the desired goal of 80%, it is important to note that percentages for both matching of codes and labels as well as overlapping time improved with each attempt. Further, results from those students who took the test more than three times demonstrated increasingly more proficiency with repeated attempts (see Table 1) attaining moderately-high to high levels. On average, those who submitted four or more attempts (n = 17) averaged 83% matching codes and labels and 75% overlapping time.

The number of errors was tracked for each of the three attempts to ascertain which of the codes and labels may have been problematic. Since error rates (≥ 40%) were consistently found in four competencies, these results provide some valuable implications for future class instruction to alleviate these high error rates in future students (Dye, 2007).
Significant within-group differences between trials (attempts 1 through 3) reveal; significant improvements (omnibus ANOVA $F(2, 58) = 24.492, p < .001$ (see Table 1), TUKEY HSD post hoc comparisons (see Table 2). Improvements were likely due to the feedback given upon submission of each attempt. While not surprising, it is an indication that within a short time and coupled with specific feedback, coding may be a skill within the learning capabilities of students.

Class membership also influenced coding ability with Section A outperforming Section B on all three attempts with significant between-class differences in matching codes being noted on the first and third attempts and in overlapping time on the third attempt (see Table 1). As mentioned earlier, the primary investigator attended and observed both sections making note of differences between sections that potentially might affect student coding proficiency results. While there are some differences in points of emphasis each favors, the researcher noted a remarkable amount of alignment between these two instructors, content, and how they teach this class. However, some specific differences that may have influenced the outcome on the CPT include: Section A had approximately 60 minutes more in-class time on the CPT with the instructor present to clarify codes and labels only—but did so without influencing the way students coded the video; instructor A gave a rationale for using DV, “seeing is better than remembering”; instructor A made a formal introduction of the primary researcher who attended class whereas instructor B only introduced him when it was time to get signed consent for participation in the study; instructor A used the actual CPT test video whereas instructor B did not; instructor A used attentional cueing (e.g., “If you noticed the teacher making a purposeful movement toward a student, it is likely that she has noticed non-compliant behavior and is engaging in the discipline plan.”).

During follow-up interviews, the instructors attributed between class differences to the extra lab time (two hours for Section A vs. one hour for Section B) as it afforded students the opportunity to ask questions. Further, it was noted that Section A was informed that students should attempt to attain an 80% agreement rate on the CPT; Section B was told to take the test no more than three times and to do their best.
Table 1
Coding Proficiency Test Results

<table>
<thead>
<tr>
<th>Attempt</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>≥ 4th</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M(%)</td>
<td>SD</td>
<td>ES</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>.49</td>
<td>.18</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>.55</td>
<td>.13</td>
<td>--</td>
</tr>
<tr>
<td>Class A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>.63**</td>
<td>.16</td>
<td>1.31†††</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>.63</td>
<td>.08</td>
<td>1.09†††</td>
</tr>
<tr>
<td>Class B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>.43</td>
<td>.15</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>.51</td>
<td>.12</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: significant between class differences in matching codes on attempts 1 and 3 and in overlapping time on attempt 3, $p \leq .01$. Effect size ($ES=(M1-M2)/SD_{pooled}$), †=small effect size (<0.40), ††=moderate effect size (0.41-0.70), and †††=large effect size (>0.70) indicating the magnitude of the repeated measures effect.
Table 2
Pairwise Comparisons of Proficiency Tests Attempts

<table>
<thead>
<tr>
<th>Attempts compared</th>
<th>Mean Difference</th>
<th>SE</th>
<th>p</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2</td>
<td>-8.48</td>
<td>2.948</td>
<td>.024</td>
<td>.44††</td>
</tr>
<tr>
<td>2 3</td>
<td>-9.11</td>
<td>1.939</td>
<td>.001</td>
<td>.31†</td>
</tr>
<tr>
<td>1 3</td>
<td>-17.593</td>
<td>2.552</td>
<td>.001</td>
<td>1.25†††</td>
</tr>
</tbody>
</table>

Note: post hoc comparisons calculated via Tukey HSD; Effect size (ES = (M₁-M₂)/SDpooled), † = small effect size (< 0.40), †† = moderate effect size (0.41-0.70), and ††† = large effect size (> 0.70) indicating the magnitude of the repeated measures effect.

Content Audit Results

The content audit results provided added understanding to the students’ coding abilities. Fifteen randomly selected instances from the solo teaching experiences of 20 students were examined for content accuracy (i.e., whether the student coded the instance using the appropriate code and label). In total, 300 codes and their labels were checked for content accuracy. Unlike the CPT results that yielded moderate to moderately-high coding proficiency, the students demonstrated a high ability to correctly identify, code, and label specific teaching behaviors found in the desired competencies list. Of the three hundred codes examined, only 26 (accuracy = 91%) were found to either be coded or labeled incorrectly. This score far exceeded the desired 80% target indicating that students can, in their own teaching videos, accurately identify, code, and label instances relating to competencies taught and practiced in class. However, an examination of the number of errors on the content audit revealed that the majority (21/26) of the errors occurred in three codes and their labels, freeze position (stopping on a signal) with students being compliant vs. non-compliant, short vs. long instructions, and step-one of the discipline plan—using positive reinforcement vs. positive praise only.

There is some consolation in the fact that all 26 of these coding errors were coded correctly, but all were simply mislabeled. For example, while the students correctly coded a freeze position, the error lay in labeling it incorrectly as all-compliant when there was actually evidence of non-compliant students. The errors seem to not be a reflection of inability to correctly identify and code the major behaviors, but rather in ability to distinguish be-
tween the more subtle differences represented in the labeling of those codes. From this we may also surmise that errors in one instance also show up as errors in other instances and may explain some of the lower CPT results by doubling up on the error count. For example, if a student were to code and label an instance as a freeze—all-compliant when it was actually a freeze—non-compliant, the CPT calculations would deduct one point for having an extra of the first and would deduct another point for missing one of the latter thus doubling the error rate.

Lastly, despite instructor efforts to remove as much ambiguity as possible in how the codes and labels were defined to the students, it proved to be a more difficult task than anticipated. Student interpretations still revealed a measure of variability in their perceptions of the coding system and how to implement it. The following are some examples that illustrate this ambiguity. Some were unsure of the appropriate length of coding an instance and were instructed to include 1-2 seconds before and after the coded instance to assure full coverage. With many clips lasting only 10-15 seconds, these 1-2 seconds could have a large influence on the overlapping CPT results. Many students realized that few teaching behaviors are executed in isolation (as would be suggested by the list of desired competencies) and were uncertain how to code simultaneous behaviors. Some were unclear if they should code instruction to individuals or only instruction to the entire class. Also, the master teacher on the CPT did not perform the competency exactly as written on the competency list, and some wondered if they should code it. Delineating between short and long instructions was occasionally ambiguous. It was also uncertain as to whether multiple positive reinforcements were to be coded separately or as a single instance when they occurred together. Overall, distinguishing between correct labels (e.g., all-compliant vs. non-compliant) was the most common difficulty. There are others, but it is important to view these difficulties in light of the fact that within a relatively short time and with practice the students’ abilities to code and label accurately improved.

**DISCUSSION**

This study was conducted to assess the ability of pre-service teachers to reliably code DV teaching episodes with the hope that if they could do so, students could become a valuable source of timely and accurate feedback during their own practice-teaching efforts. The present researchers used mixed methods to explore the level of agreement of novice coders with
expert coders as well as to identify influences that either contributed to or detracted from their ability to do so. This paper initially proposed that DV analysis could increase the amount and accuracy of feedback that students receive over and above that of traditional observation/debriefing or reflection practices. It also proposed that DV analysis might be a means of reducing such things as memory decay, disparate perceptions, and communicative breakdowns. Some of these propositions were supported in this study while others were not. We will first discuss the implications of the findings with respect to coding reliability and then whether DV analysis can be considered a viable means to augment feedback as well as overcome some of the weaknesses that hamper such practices.

Reliability of Pre-Service Teachers’ Coding

Results indicate that students, within a short time (one to two hours), can achieve moderate to high levels of agreement with expert coders when respectively coding a master teacher or themselves. Further, it is clear that with continued practice, students’ coding abilities continue to rise. This is notable when considering the compressed nature of this course. Future research should examine coding proficiency in programs where pre-service teacher development occurs over successive semesters to assess if coding accuracy continues to rise and stabilize at high levels of proficiency. While these results are encouraging, certain difficulties must be addressed if DV analysis is to become a viable source of feedback.

For example, despite efforts to control for extraneous sources of variance (e.g., poorly matched competency definitions and video clip samples) in this study, it became clear that a degree of ambiguity existed in the coding process, which undoubtedly had an effect on the interrater reliability scores. The majority of the ambiguity is attributed to the instructors’ belief that the list of desired competencies and the coding system was unambiguous. This erroneous assumption revealed the need for ongoing refinement of how it is presented and applied in future classrooms, particularly in compiling video training clips that match precisely the content of the competency list. When asked to explain possible reasons for the apparent ambiguity faced by the students, the instructors commented on discrepancies between novice and expert coding in what they see and also that “behavioral chains” do not easily segment themselves into the neatly packaged set of behaviors in the competency lists. The instructors concluded that creating training materials that precisely match the desired competencies is paramount and could remove
much of the ambiguity students experienced. Their recommendation for
teacher educators include (a) clearly defining context-specific desired com-
petency outcomes suited to the constraints of the class, (b) creating video
clips to match desired competencies, (c) well defined scoring rubrics made
available to students up front, (d) in-class instruction and providing feed-
back during the Studiocode learning phase, and (e) refining the approach
and materials as needed.

The content audit gave some additional insight as to where the majority
of the errors lay—in the labeling and not the application of the codes. Cod-
ing accuracy is seemingly linked to one’s ability to distinguish not between
codes but between the more subtle labels applied to those codes. In the fu-
ture, this ambiguity can be greatly alleviated by providing training that uses
footage precisely matched to the list of desired competencies represented in
the coding/labeling system. However, despite these difficulties, and in light
of the moderate to moderately high reliability outcomes, the use of DV self-
analysis shows great potential as a source of accurate and valuable feed-
back.

Addressing Increased Feedback, Memory Decay, Disparate Perceptions, and
Communicative Breakdowns

In the past, video analysis was a cumbersome and time-consuming ef-
fort that made it difficult to justify its use in teacher education classes. This
study provides evidence that within a short amount of time (one to two
hours) students can be taught to use Studiocode software and can be trained
to code using a predefined set of codes and labels with increasing accuracy
as they continue to practice. In this study, out-of-class DV evaluations of
one’s own teaching were completed in less than 90 minutes per week. Thus,
the “time sink” of yesteryear appears to no longer be a deterrent to using
DV analysis in teacher preparation programs. As a result of this modicum of
time invested, students felt that they had indeed received far more feedback
than had they not analyzed their weekly teaching episodes. One student re-
ported the following:

“The value of watching ourselves teach the PE lessons in the
schools is much greater than just trying to remember how our les-
sions went that day. It really helps to watch ourselves on the video.
We can see what we are doing wrong, and then also we can see
some kids that are misbehaving that we didn’t catch during the
lesson. It is an accurate account of how we did, and it really helps us to see what we need to work on for the next week. If we just remembered back on the lesson, then I think a lot of things would be forgotten. It really helps to watch ourselves.”

As indicated by this student, perceptions were often mistaken and were more accurate from the use of DV analysis. The most common student theme from all student responses with respect to the value of using DV analysis was that students had no idea that certain things were going on in their lessons. “Stuff,” one student report, “was happening right in front of me and I did not even see it. But there it was right in the tape.” Yet another student recounted her experience with DV analysis as follows:

“I definitely think there’s A LOT of value in having the technology available to be able to see ourselves teach. There’s only so much you can say and evaluate on when you’re reflecting on how you THINK you did. If any students who are being taught to be teachers are like me, you tend to not see things for the way they are. What I mean is that for myself I found that I either thought I was really, really horrible and didn’t realize the good things that I did, or I thought I was pretty darn good and didn’t realize the techniques I needed to work on.”

Studiocode also provides a transcription tool for students to write comments to accompany each of the coded instances. The text appears at the bottom of the clip. It can be used to transcribe dialogue verbatim, or write personal analyses, reflections or improvement plans. This feature allows for video-aided reflection while reducing memory decay and disparate perceptions.

CONCLUSIONS

Despite the fact that students failed to meet the expected level (i.e., 80%) of agreement and especially considering the student responses, the instructors remain positive about the value of DV self-analysis in the course. They maintain that this iterative process of analysis, filtering, and selection of clips and subsequent planning of lessons to address deficiencies includes much of the value of previous assessment/feedback practices discussed earlier while avoiding their shortcomings.
“Ultimately,” one instructor said, “What we really want to know is how well a candidate can teach. In the past we have graded using observation checklists or on how well the student could reflect on their teaching but not how well they taught. Digital video allows students to view, reflect upon, and make value judgments of their own performances and address weaknesses. Then teacher educators can truly evaluate how well they can perform the desired competencies.”

The iterative process of student driven DV analysis using Studiocode software in combination with the final instructor evaluation of actual teaching performance provides an assessment that is both formative and summative in nature. A future research question is how well students can make value assessments of their teaching performances. It seems evident, from the results of this study, that students can become capable analyzers, but it remains unclear if they can accurately assess the quality of the performance. If students can make accurate value assessments, they will likely be more able to address deficiencies. This underscores the importance of making available to students a clearly defined set of desired competencies coupled with matching video clips as well as a clear scoring rubric.

Lastly, the authors explore the notion that DV analysis may provide a means to aid in (a) identifying and mastering a new skill set necessary for self-analysis resulting in increasing one’s “noticing” abilities and (b) the aligning of inert PCK and authentic practice. While the results of this study are encouraging on both accounts, much remains to be done to examine if DV analysis can indeed “shorten the learning curve” for pre-service teachers. In particular, while it is important to find that students can become adequate self-analyzers, the greater question remains—does DV self-analysis, especially if used over successive semesters of their preparation, lead to better teaching performance? Future experimental research should compare the addition of DV self-analysis and its effects on teaching performance over observation and reflection alone.
References


Figure 1 Studiocode® digital video analysis software allows users to view and code teaching episodes using the following windows: upper left = video window; upper right = transcription window; lower right = time line containing coded/labeled instances; lower left = code input window.
### DESIRED COMPETENCIES USED IN THIS TWO-CREDIT HOUR, ELEMENTARY PHYSICAL EDUCATION METHODS CLASS.

**Elementary Physical Education Desired Competencies**

<table>
<thead>
<tr>
<th>Desired Competency</th>
<th>Critical elements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Freeze positions</strong></td>
<td></td>
</tr>
</tbody>
</table>
| i. Freeze - All Compliant | 1. Teacher Talk (rehearsed, clear, concise, seldom uses idiomatic phrases “umm, ok, etc.”)  
2. Teacher on the perimeter  
3. Model the freeze  
4. Quietly scan side to side  
5. Use positive praise; specific to all parts of the freeze position to more than one student  
6. Stand them tall right before giving the next instruction |
| ii. Freeze - 1-4 non-compliant | 1. Teacher Talk  
2. All of the above components (freeze-all compliant)  
3. Use of all appropriate steps in the discipline plan (positive reinforcement, proximity, warning, etc.) to deal with the non-compliance |
| **b. Transitions involving Equipment** | |
| i. Moving Equipment onto the floor | 1. Teacher Talk  
2. Equipment is set up spaced around the perimeter of the floor  
3. Use When before What  
4. Say “Watch me first!”  
5. Model how to get the equipment and find your own space. Then model the first activity they should do (transitional activity)  
6. Completed in a timely manner (after the go, no more than 10-15 seconds until all students are engaged) |
### ii. Moving Equipment off the floor

<table>
<thead>
<tr>
<th>Step</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher Talk</td>
</tr>
<tr>
<td>2.</td>
<td>Reverse the above procedure</td>
</tr>
<tr>
<td>3.</td>
<td>Emphasize and model carrying the equipment (not playing with it) and placing it down (not throwing or rolling it)</td>
</tr>
<tr>
<td>4.</td>
<td>Give students a transitional activity for after the equipment is put away</td>
</tr>
<tr>
<td>5.</td>
<td>Complete in a timely manner (after the go, no more than 10-15 seconds until all students are engaged)</td>
</tr>
</tbody>
</table>

### c. Transitions moving students

#### i. Making partners (toe to toe)

<table>
<thead>
<tr>
<th>Step</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher Talk</td>
</tr>
<tr>
<td>2.</td>
<td>Freeze the class first, scan, and reinforce</td>
</tr>
<tr>
<td>3.</td>
<td>Use “When” before “What,” “Class, when I say toe-to-toe, find the person closest to you and stand like this (model this if the first few times taught): Toe-to-toe”</td>
</tr>
<tr>
<td>4.</td>
<td>After the “Go”, remind those that don’t have partners to hustle to the middle, raise their hands, and find a partner</td>
</tr>
<tr>
<td>5.</td>
<td>Complete in a timely manner (10-15 seconds after the go)</td>
</tr>
</tbody>
</table>

#### ii. Splitting the class into two teams

<table>
<thead>
<tr>
<th>Step</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher Talk</td>
</tr>
<tr>
<td>2.</td>
<td>Use toe-to-toe as described above</td>
</tr>
<tr>
<td>3.</td>
<td>Have one person sit down</td>
</tr>
<tr>
<td>4.</td>
<td>Model where they should go and how they should stand, then move the standers first while leaving the sitters in place</td>
</tr>
<tr>
<td>5.</td>
<td>Once the standers are in place, move the sitters to where they should go</td>
</tr>
<tr>
<td>6.</td>
<td>Completed in a timely manner (under 30 seconds)</td>
</tr>
</tbody>
</table>
### iii. Making groups (whistle mixer)

1. **Teacher Talk**
2. Freeze the class first, scan, and reinforce
3. Use *When* before *What*, “Class, show me with your fingers how many times I blow my whistle (model this with your own fingers). When I say go, make a group with that many people and sit down in a straight line facing me. Go.”
4. After the Go, remind those that don’t have a full group to hustle to the middle, raise their hands, and find a full group
5. Complete in a timely manner (10-15 seconds after the go)

### d. Instructional episodes

#### i. Using “When” before “What”

1. **Teacher Talk**
2. Use the phrase “When I say Go, I would like you to . . . Go” (do not use the whistle to start)
3. Model the activity you want them to do during your explanation and before you say, “Go”
4. Beware of the creep! If students begin to move before you say “Go,” you may need to stop the instructions and do a move and freeze, reminding them about listening to all the instructions, then repeat, “When I say Go, . . .”

#### ii. Giving short instructions (< 1 minute)

1. **Teacher Talk**
2. Freeze the class, scan, reinforce and stand them tall
3. Give the instructions using concise teacher talk, direct and to the point. Do not compete with talking students. Model the instructions
4. When appropriate, use the phrase, “When I say Go, . . .” to get the class started again
5. After the “Go” scan, check cards, if necessary, and reengage with the class, working the crowd
### Giving long instructions with students sitting down (> 1 minute)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher Talk</td>
</tr>
<tr>
<td>2.</td>
<td>Freeze the class, scan, reinforce and sit them down where they are</td>
</tr>
<tr>
<td>3.</td>
<td>Give the instructions using concise teacher talk, direct and to the point. Do not compete with talking students. Model the instructions</td>
</tr>
<tr>
<td>4.</td>
<td>When appropriate use the phrase “When I say Go, . . .” to get the class started again</td>
</tr>
<tr>
<td>5.</td>
<td>After the “Go” scan, check cards if necessary, and reengage with the class working the crowd</td>
</tr>
</tbody>
</table>

### Using the Discipline Plan

#### i. Step 1: Positive reinforcement

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher Talk</td>
</tr>
<tr>
<td>2.</td>
<td>Use positive reinforcement to change non-compliant behavior</td>
</tr>
<tr>
<td>3.</td>
<td>Clearly identify to the class a compliant student, use their name, and positively reinforce the specific behavior of the non-compliant student</td>
</tr>
</tbody>
</table>

#### ii. Step 2: Teacher Proximity

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher Talk</td>
</tr>
<tr>
<td>2.</td>
<td>Continue teaching/reinforcing while you move next to the problem student(s)</td>
</tr>
<tr>
<td>3.</td>
<td>Make note if the behavior is corrected; if needed, be ready to give a warning</td>
</tr>
</tbody>
</table>

#### iii. Step 3: Quiet Warning

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher Talk</td>
</tr>
<tr>
<td>2.</td>
<td>Reengage the class</td>
</tr>
<tr>
<td>3.</td>
<td>Issue the warning, state the misbehavior, state the rule, state the consequence. For example, “Todd, you were talking while I was giving instructions, that’s rules number 1 and 2; this is your warning, next time you’ll go to time out.”</td>
</tr>
<tr>
<td>4.</td>
<td>Walk away and reengage with the class</td>
</tr>
</tbody>
</table>
| iv. Step 4: First Time Out | 1. Teacher Talk  
2. Reengage the class  
3. Issue the Time Out, state the misbehavior, state the rule, tell them to go to time out. For example, “Todd, you pushed your classmate, keeping your hands to yourself is part of rule number 2, please go to time out. You may return when you are ready to behave properly.”  
4. Walk away and reengage with the class and give some positives. Take note if the student went to time out |
|--------------------------|---------------------------------------------------------|
| v. Step 5: Second Time Out | 1. Teacher Talk  
2. Reengage the class  
3. Issue the Time Out, state the misbehavior, state the rule, tell them to go to time out, remain until the end of class, and then come see the teacher. For example, “Todd, this is the third time I have had to talk to you today. You were not following the instructions, this is rule number 1, please go to time out, stay until the end of class, and then come see me.”  
4. Walk away and reengage with the class and give some positives. Take note if the student went to time out |