2007-07-18

Reliability of Pre-Service Teachers Coding of Teaching Videos Using Video-Annotation Tools

Brigham R. Dye
Brigham Young University - Provo

RELIABILITY OF PRE-SERVICE TEACHERS' CODING OF TEACHING VIDEOS USING A VIDEO-ANALYSIS TOOL

by

Brigham R. Dye

A thesis submitted to the faculty of

Brigham Young University

In partial fulfillment of the requirements for the degree of

Master of Science

Department of Instructional Psychology & Technology

Brigham Young University

June 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

Of a thesis submitted by

Brigham R. Dye

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Charles R. Graham, Chair
Date	Keven A. Prusak
Date	Russell T. Osguthorpe

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Brigham R. Dye in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_____ _____

Date

Charles R. Graham
Chair, Graduate Committee

Accepted for the Department

_____

Andrew S. Gibbons
Department Chair

Accepted for the College

_____

K. Richard Young
Dean, College of Education

ABSTRACT

RELIABILITY OF PRE-SERVICE TEACHERS' CODING OF TEACHING VIDEOS USING A VIDEO-ANALYSIS TOOL

Brigham R. Dye

Department of Instructional Psychology and Technology

Master of Science

Teacher education programs that aspire to helping pre-service teachers develop expertise must help students engage in deliberate practice along dimensions of teaching expertise. However, field teaching experiences often lack the quantity and quality of feedback that is needed to help students engage in meaningful teaching practice. The limited availability of supervising teachers makes it difficult to personally observe and evaluate each student teacher's field teaching performances. Furthermore, when a supervising teacher debriefs such an observation, the supervising teacher and student may struggle to communicate meaningfully about the teaching performance. This is because the student teacher and supervisor often have very different perceptions of the same teaching performance. Video analysis tools show promise for improving the quality of feedback student teachers receive in their teaching performance by providing a common reference for evaluative debriefing and allowing students to generate their own feedback by coding videos of their own teaching. This study investigates the reliability of pre-service teacher coding using a video analysis tool.

This study found that students were moderately reliable coders when coding video of an expert teacher (49%-68%). However, when the reliability of student coding of their own

teaching videos was audited, students showed a high degree of accuracy (91%). These contrasting findings suggest that coding reliability scores may not be simple indicators of student understanding of the teaching competencies represented by a coding scheme. Instead, reliability scores may also be subject to the influence of extraneous factors. For example, reliability scores in this study were influenced by differences in the technical aspects of how students implemented the coding system. Furthermore, reliability scores were influenced by how coding proficiency was measured. Because this study also suggests that students can be taught to improve their coding reliability, further research may improve reliability scores—and make them a more valid reflection of student understanding of teaching competency—by training students about the technical aspects of implementing a coding system.

ACKNOWLEDGMENTS

I am indeed grateful to the many individuals who have helped make this thesis possible. This thesis would not have been possible if Keven Prusak had not opened his course and innovative pedagogy to my research efforts. He and Sue Graser pioneered the use of video analysis tools in their classes and generously allowed me to learn from their experience. I thank both of them for their willingness to invite me into their classrooms and to help me make sense of this study each step of the way. I am particularly grateful to Charles Graham for his mentoring. He has shown genuine interest in my learning and professional development in all of our interactions. This thesis is the result of several lengthy conversations where he helped me grapple with challenging concepts. I also thank Russell Osguthorpe who has been willing to make himself available whenever it was needed. I will confess I overheard him say something kind about me to another committee member at my prospectus defense, and that gave me a needed boost of confidence.

In particular, I want to thank my sweet wife, Kadee, for the support she has given me. She is the one who had to live with a husband who was writing a thesis…and she did it with great patience and love. Her expressions of confidence in me mean more than she can realize.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

Because teaching is a performance-based profession, teacher education programs integrate field-based practice with classroom-based instruction. A typical teacher education program scaffolds the process by which novice pre-service teachers gradually assume responsibility for more complex teaching events. Field experience may begin with peripheral observation of classroom interactions. Then pre-service teachers may partner with an experienced teacher in presenting a lesson. Eventually, the initiate teacher assumes responsibility for classroom instruction while the experienced teacher plays a more peripheral role as observer/evaluator.

The purpose of this transitional process in which novice teachers progress from peripheral observers to central performers is to support pre-service teachers as they develop characteristics of teaching expertise. Development of expertise requires deliberate practice along dimensions in which expert teachers show unique ability. Effective practice requires meaningful and appropriate feedback.

When pre-service teachers engage in field experiences, they typically receive feedback from a supervising teacher educator or experienced teacher. This expert feedback can help to focus the practice of pre-service teachers; however, its helpfulness is limited by the costliness of providing expert feedback and by breakdowns in the communication of feedback from experts to novices.

Video analysis tools may help overcome the obstacles to providing meaningful feedback. Recent technological innovations that take advantage of the affordances of digital video make it possible for pre-service teachers to analyze videos of their own teaching performances. This analysis involves identifying and coding events from the teaching performances as instances of key teaching principles, practices, and behaviors. Preliminary

work with video analysis in physical education (PE) teacher education at BYU indicates that pre-service teachers may be able to learn to use a clearly defined set of criteria to evaluate videos of their own teaching. If video analysis helps pre-service teachers provide themselves with reliable and valid self-evaluations, then the costs and challenges of relying mainly on external evaluation could be ameliorated. Video analysis of performance may improve the quantity and quality of feedback available to pre-service teachers, the kind of feedback needed to engage in more focused and effective practice.

However, before researchers can investigate the effects of video analysis on performance, they need to establish whether students can analyze their own videos in a way that leads to reliable and valid conclusions. The purpose of this study is to consider the extent to which pre-service teachers in a course on teaching PE can be taught to reliably apply a set of expert-defined codes to videos of their own teaching performance by using a video analysis software tool.

Problem Overview

Coursework in teacher education programs often focuses on the development of pedagogical content knowledge. Pedagogical content knowledge is the blending of content knowledge and pedagogical knowledge. It involves knowing how specific topics, problems, or questions can be organized, represented, adapted, and taught to learners (L. S. 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 knowing, this pool of pedagogical content knowledge tends to be 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 (Berliner, 1992). Performance-based practice, such as that in which pre-service teachers engage during their field experiences, helps to transform this inert, declarative knowledge into viable, performative knowledge.

Figure 1 illustrates the typical relationship between pre-service teachers' theoretical pedagogical content knowledge and their actual teaching performance as they progress along the continuum from novice teaching to greater teaching expertise. When beginning field experiences, a large portion of these pre-service teachers' actual teaching performance aligns more closely with their personal assumptions and prior educational experience than with the body of pedagogical content knowledge they have been developing through their classroom-based coursework (Winograd, 1995). Appropriate feedback can serve to highlight the misalignment of theory and practice and help pre-service teachers to begin the process of bringing their actual teaching performance and pedagogical content knowledge into convergent dialogue. By the phrase, *convergent dialogue*, I mean to suggest that not only does theoretical knowledge influence and change teaching performance but also that the teaching performance can inform and shape the pre-service teacher's theoretical knowledge. Through iterative cycles of practice and feedback, the pre-service teachers can develop greater teaching expertise as characterized by increased alignment of practice with theoretical knowledge. Simultaneously, their pool of pedagogical content knowledge increases and/or is dynamically reorganized.

In actuality, the relationship between classroom-based development of pedagogical content knowledge and field-based development of performative knowledge is not as linear as I have described it. Strong teacher education programs work to integrate the two types of

knowledge and learning environments. For example, classroom use of case-based studies can involve pre-service teachers in authentic practice with the kind of reasoning teachers engage in, as if they were actually in a field experience. However, the distinction is useful for understanding the roles of feedback and of field experiences.

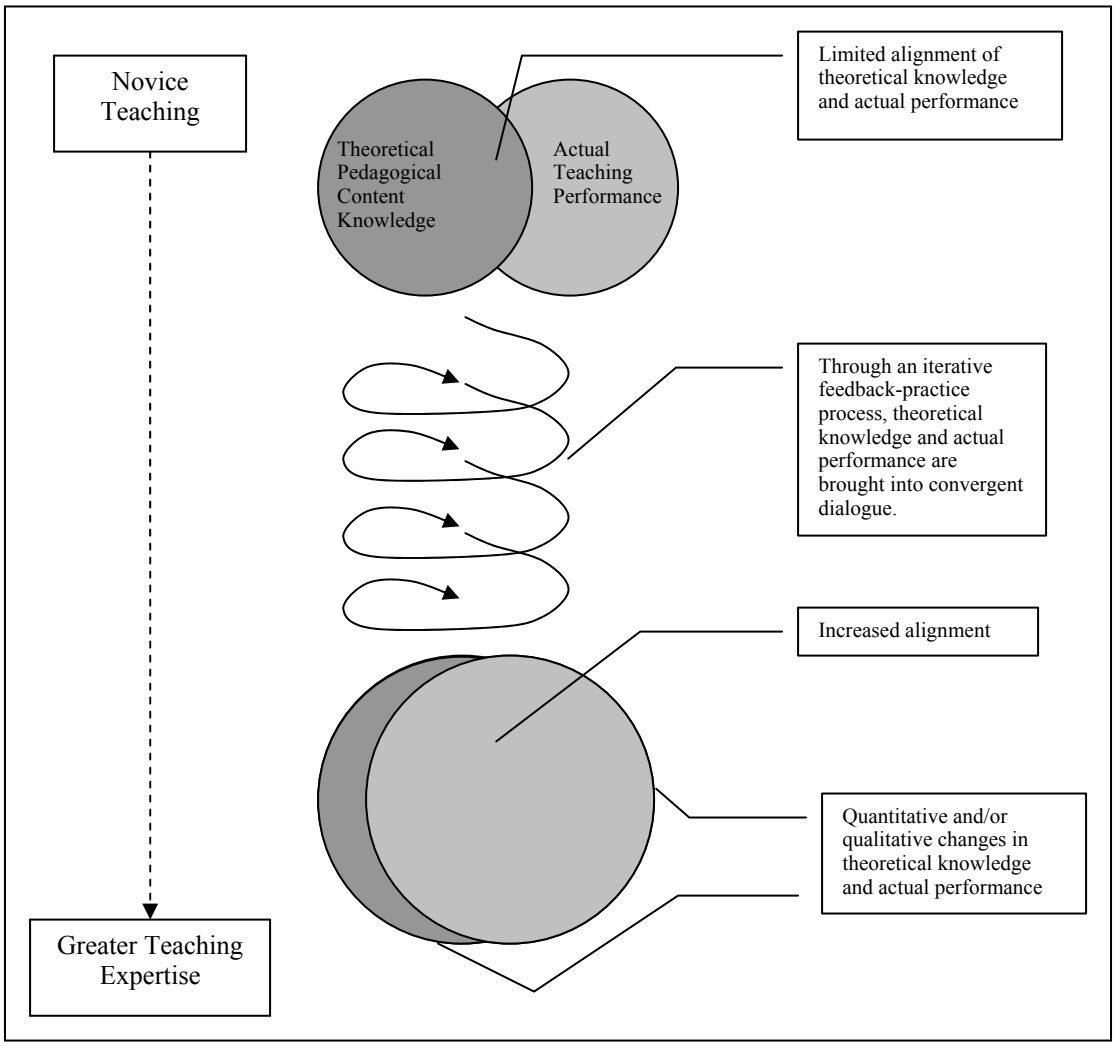

Figure 1. Development of teaching expertise through the progressive alignment of theoretical knowledge and teaching performance.

The iterative feedback-practice process that pushes teachers towards increased teaching expertise requires enough accurate and meaningful feedback to focus practice on the dimensions of the teaching performance that characterize teaching expertise. The impact of

feedback in teacher education is limited by perceptual and communicative breakdowns. A common practice in teacher education is for an external evaluator, a mentor teacher or a university supervisor, to observe the pre-service teacher giving a lesson and then provide evaluative feedback following the lesson. One of the problems with this feedback is that both the external evaluator and the pre-service teacher have their own unique perceptions of the teaching performance. Consequently, when the pre-service teacher and the evaluator attempt to debrief the teaching performance their ability to communicate is limited by the degree to which they share a common perception of the teaching performance.

Figure 2 shows some of the influences that shape perceptions of the teaching performance. For the performer, his/her experience of the performance is shaped by the temporal nature of the performance, by the lack of an experiential background, and by his/her own limited processing capacity. For the novice, the sheer complexity of the performance may overwhelm his/her ability to access and make use of his or her pedagogical content knowledge. For example, a novice teacher may focus on the logistics of delivering a lesson plan while failing to attend to the experiences of individual students in the class. The degree to which novices have developed automaticity in many of the complex, though more trivial, prerequisite skills of performance will affect their ability to notice opportunities to apply their advanced pedagogical content knowledge. Following the performance, the performer's recollection of the performance will be limited by the ways in which his/her attention was focused while performing.

Although the external evaluator is not subject to the same processing constraints as the performer, his/her perspective is nonetheless also limited by how he/she selectively focuses his/her attention. By analogy, this is much like the sampling problem in experimental studies. Because the performance is temporal in nature, the external

evaluator can only sample a portion of the many teaching/learning phenomena that are occurring in the classroom. This sampling is not likely to be random, but rather is shaped by such things as the evaluator's expectations, values, interests, and prior experience.

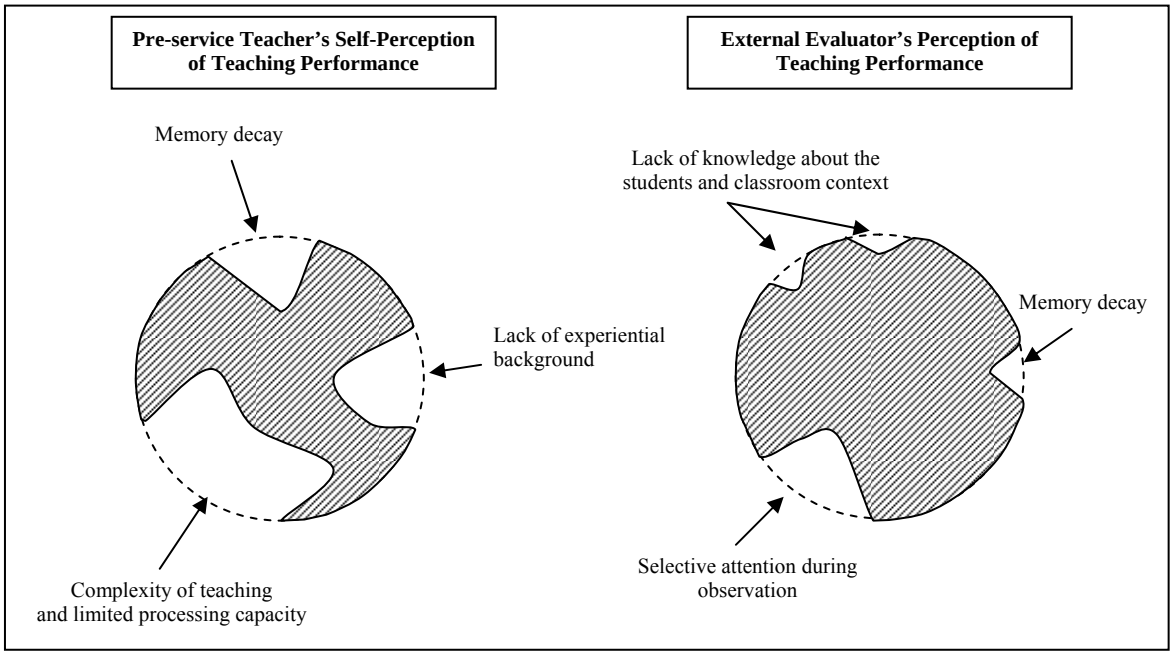

Figure 2. Factors that influence how pre-service teachers and evaluators perceive a teaching performance.

The evaluator's perceptions are also limited by a lack of familiarity with the context of the lesson being observed. Unlike the pre-service teacher who may be familiar with the needs or history of specific students, the external evaluator is not likely to be intimately aware of the context for teacher behaviors. Likewise, the external evaluator is not likely to be aware of the entire class's collective history, routines, and needs. This lack of contextual awareness can produce gaps in the evaluator's perceptual understanding of the teaching performance (West, 2007).

The limitations of memory affect the perceptions of both the evaluator and the pre-service teacher. By the time these individuals meet to debrief the lesson, it is likely that memory decay has taken its toll on the accuracy of perceptions.

During the debriefing session during which the evaluator provides feedback to the pre-service teacher, there are likely to be communication breakdowns due to their differing perceptions of the teaching performance. Research provides strong evidence that feedback messages are complex and require active interpretation by students (Nicol, & Macfarlane-Dick, 2006). One of the communicative filters that affect this interpretation process results from the lack of a shared perception. When an external evaluator shares his/her evaluation of the performance with a pre-service teacher, the pre-service teacher interprets the message in terms of his/her own perception of the experience of the performance.

Figure 3 shows two possible scenarios for how perceptions might influence the communication of feedback. In the first scenario, the evaluator and the pre-service teacher discuss an aspect of the teaching performance for which they have similar perceptions. In this case, they share enough common ground to be able to meaningfully communicate about the teaching performance. In the second scenario, however, the evaluator and pre-service teacher discuss an aspect of the teaching performance for which they have widely differing perceptions. In this case, they are less likely to be able to communicate effectively about the teaching performance. Furthermore, neither of the individuals has complete information about the actual teaching performance and so feedback is more likely to be inaccurate.

Video analysis software tools provide an opportunity to reduce perceptual and communicative breakdowns in the evaluative feedback that pre-service teachers receive. First, video analysis tools make the temporality of the performance dynamic and flexible. Rather than having to rely on one's singular experience of and limited memory of the performance, those using video analysis tools can engage in multiple viewings of the

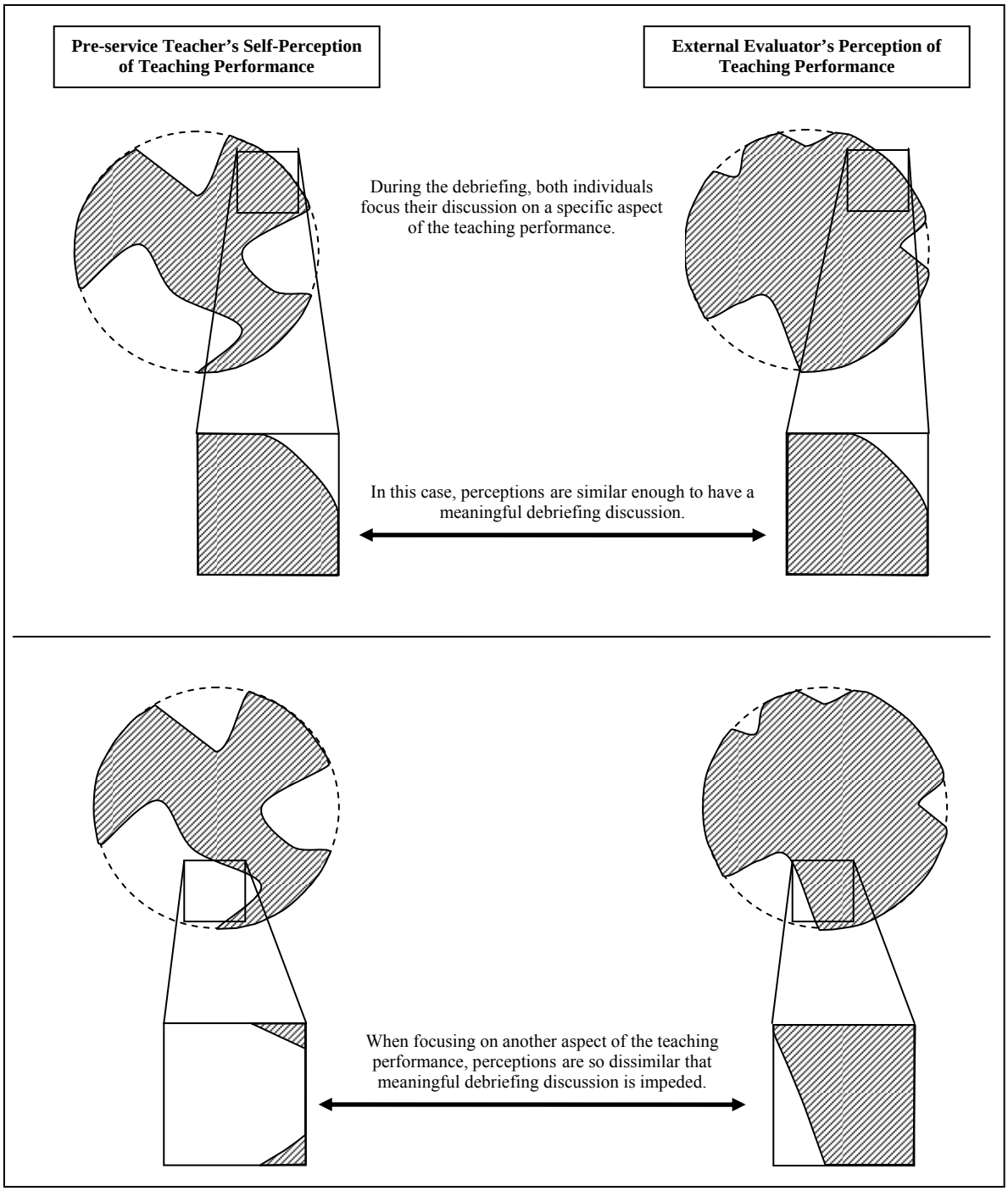

Figure 3. Communication breakdowns during a feedback session.

performance. This allows the user to view the performance from various perspectives and to re-examine portions of the performance that seem unclear or complex. By referencing a high fidelity representation of the performance (i.e., a video recording of the performance) the performer and external evaluator can both make inquiries about their own mental representations of the performance.

Second, video analysis tools provide a shared representation of the performance around which the performer and external evaluator can communicate. Because video analysis software allows the user to index the video of the performance, communication can be anchored in specific events. Rather than speak in vague generalities about teaching performance, the external evaluator and performer can direct each other's attention to particular events to illustrate their ideas.

In addition to these benefits, video analysis tools can also reduce the costliness of providing expert external evaluation. When an expert evaluator is supervising many pre-service teachers, it is unrealistic to observe every teaching performance from every pre-service teacher. Evaluative feedback thus tends to assume somewhat of a scattershot approach. Classroom circumstances during the evaluator's visit may not occasion a specific teaching-learning interaction for which evaluative feedback is needed. However, with video analysis software, the pre-service teacher can pre-index videos of his/her own performance. The indexed video then provides an opportunity for focused expert feedback.

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, during the performance the external evaluator is not engaged in performing and so 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 functions of the external evaluator.

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 of which they may have been unaware while in the act of performing. Thus, the first function of an external evaluator can be fulfilled by the performer.

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. In a course on physical education for elementary school teachers at Brigham Young University, a video analysis tool scaffolds this noticing process by providing a set of clearly defined codes which have been extracted from the literature on relevant pedagogical content knowledge. Pre-service teachers are explicitly directed to watch for instances of these key dimensions of expert performance in their own videos. Prior to coding their own videos, however, pre-service teachers are trained in the meaning of and application of these codes. They practice applying the codes to videos of expert PE teachers. Thus trained, pre-service teachers are then asked to code their own videos.

Research Questions

What remains to be determined is the degree to which pre-service teachers can arrive at reliable and valid evaluative conclusions based on their analyses of their own videos.

This study addresses this problem by answering the following questions:

1. How reliable are pre-service teachers' analyses of videos of expert teachers when compared to the analyses of experts?
2. How reliable are pre-service teachers' analyses of videos of their own teaching performance?

Only if we first know that students can be trained with video analysis software to reliably evaluate their own teaching can we then consider whether students' self-evaluations lead to improved practice. The findings of this study could potentially lead to improved feedback for pre-service teachers during their field experiences. Improved feedback could then lead to engagement in more deliberative practice and ultimately better teaching.

Chapter 2: Review of the Literature

Teaching Expertise

By comparing novice and expert teachers, researchers in teacher education have identified a number of characteristics that describe teaching expertise. Expert teachers tend 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, 2007). One of the most important abilities of an expert teacher is the ability to notice key features and patterns in a teaching situation that often escape the novice teacher (Bransford, 2006). Experts have fast and accurate pattern-recognition capabilities and develop more complex, specific, and strategic representations of classroom situations and problems than novices (Berliner, 1992; Hogan, 2003). Not only do experts perceive more meaningful patterns in teaching/learning situations, but the types of things various experts notice tend to converge (Berliner, 1992). This finding suggests the existence of a distinct professional body of teacher knowledge, even if that body of knowledge is as of yet ill-defined (Hiebert, 2002).

The unique perceptual and representational abilities of expert teachers are essential for decision-making in a fast-paced, complex classroom environment. In one study, the classroom properties of *simultaneity*, *multidimensionality*, and *immediacy* were simulated to compare novice and expert teacher's perceptual capacities (Sabers, 1991). Simultaneity refers to the multitude of interactions among students and the teacher that take place at the same time in the classroom; multidimensionality has to do with the large number of events and tasks going on in a classroom, the layered-ness of teaching; and

immediacy refers to the rapidity of responding to all of these simultaneous interactions. In the study, experts and novices simultaneously viewed three television monitors showing three sectors of the same classroom. Researchers found that expert teachers were better able to distribute their attention across the three monitors and to interpret the instructional strategies used, formulate hypotheses to explain student behaviors, and suggest solution strategies. Whereas both experts and novices were equally capable of remembering trivial classroom events, experts were better at remembering meaningful classroom interactions.

Not only is teaching expertise enacted in a simultaneous, multidimensional, and immediate environment, but teaching expertise also seems to develop in response to this environment (Hiebert, 2002). Merseth argues that teacher knowledge is dynamically emergent, embedded in the context of teaching:

> Recent research on teacher thinking has broadened the conceptualization of the teacher from the one who operates with a narrow set of prescribed theories of propositions to one who defines his or her knowledge as situation-specific, context dependent, and ever emerging.... Teacher action derives from induction from multiple experiences, not deduction from theoretical principles. (Nemirovsky, 2005, p. 724)

Thus expertise in teacher thinking is situated, complex, and evolving. Furthermore, it requires not only content and pedagogical knowledge but also the contextualized blend of pedagogical content knowledge (Hiebert, 2002).

Research on teaching expertise stems from the foundational work on general expertise conducted by Chase and Simon. Their seminal study considered the differences between novice and expert chess players (Chase, 1973). Their finding that experts processed

information in larger perceptual chunks than novices has been extended to the other fields, including physical education (Chen, 2002).

Expertise has also been considered in the field of anthropology in which Goodwin (1994) has explored the concept of professional vision. Goodwin argues that professional vision describes the ability of professionals to interpret the messy world of practice through the categories unique to a particular community of inquirers/practitioners. This is accomplished through processes of coding, highlighting, and visual representation. Coding applies a set of categories to define the world in terms that are relevant/meaningful to the work of a profession. Highlighting involves selectively making specific phenomena from a complex perceptual field salient and thus relevant to the professional's work. Finally, professionals produce or articulate a material representation of the world in a way that facilitates communication and reasoning about the professional's work. Goodwin argues that professional vision is socially constructed:

> The ability to see a meaningful event is not a transparent, psychological process but instead a socially situated activity accomplished through the deployment of a range of historically constituted discursive practices. [...] All vision is perspectival and lodged within endogenous communities of practice. (Goodwin, 1994, p. 606)

Among the theorists involved in describing expertise or professional knowledge, perhaps the best known is Donald Schon. Schon's (1983) trans-disciplinary inquiry into expertise, or what he calls reflection-in-action, explores the ways in which professionals not only reflect on their practice but also reflect in practice. By analyzing the ways in which practitioners engage in dialogue with a specific context, Schon establishes an epistemology of practice that bridges the traditional theory-practice dichotomy.

Teacher educators have been attracted to Schon's theory of reflection-in-practice because it describes the kind of tacit knowing that is essential to good teaching (Brophy, 2004). Schon's explanation of how a professional can intuitively draw upon a bank of experience and tacit knowledge to flexibly engage with a particular contextualized problem describes the types of abilities that mark teacher expertise.

Video Affordances for Developing Teaching Expertise

Teacher educators have explored a variety of instructional approaches to support pre-service teachers in developing the perceptual, representational, and reflective abilities associated with teacher expertise. Because of the situated nature of teacher thinking, teacher educators have been particularly interested in the use of video to contextualize instructional activities. Researchers on the use of video in teacher education frequently explain this preference in terms of grounded theory and situated learning (Brophy, 2004). The Cognition and Technology Group at Vanderbilt (1990) has coined the term "anchored instruction" to describe their use of video to present richly contextualized instructional problems to students. During the past two decades the use of video anchors for teacher education has proliferated.

Video provides a number of instructional affordances that support the development of teacher expertise (Brophy, 2004; Chan, 2005). Video provides access to authentic, concrete teaching episodes that otherwise might not be available to pre-service teachers. Video provides a much richer representation of teaching/learning environments that is open to inquiry from multiple perspectives and through multiple modalities. Video provides an opportunity for the development of pattern recognition skills. Video also provides a shared experience-base that facilitates communication. Furthermore, digital video changes the temporality (i.e. linearity) of teaching episodes, thus allowing for dynamic and intentional

reviewing. Thus teacher educators can systematically expose pre-service teachers to expert models and specific instructional strategies.

The affordances of video for teacher education hold great promise. Karppinen (2005) argues that through these affordances, video can be used as a constructivist pedagogical tool because it provides for at least six of the characteristics of a meaningful learning experience: active, constructive and individual, collaborative and conversational, contextual, guided, and emotionally motivating. Hiebert (2002) suggests that video technology may make it possible to establish a professional knowledge base for teaching similar to that which has been established in other professions such as law or business.

Despite the promise of video for instruction, the instructional power of video does have a major limitation. Research commonly reports that simply viewing video is not enough to develop teaching expertise (Brophy, 2004). Novice teachers in particular do not get much benefit out of self-guided video inquiry. Scaffolding is necessary to direct the attention and efforts of viewers to relevant and important aspects of the video.

Scaffolding Reflection on Videos of Other Teachers

To help pre-service teachers take advantage of the instructional affordances of video, teacher educators have developed a wide range of video tools and applications that scaffold the learners' use of video. Here I will discuss the use of video for reflecting on others' teaching. I have distinguished between using video to reflect on others' teaching performance versus using video to reflect on one's own teaching performance. If teacher knowledge is as situated as is claimed in the literature on teaching expertise, then certainly the context of who is the main performer in an instructional example is relevant to the development and transfer of teacher knowledge. In other words, there is a potential transfer gap between learning about others' teaching and learning about one's own teaching

performance. This gap is the underlying reason behind the use of clinical or field experiences in teacher education programs.

On one end of the spectrum we find instructional uses of video for teacher education that employ "soft" instructional scaffolding. Pailliotet (1995) describes a method for helping students observe and analyze video of teaching called *deep viewing*. This method is adapted from several disciplines including qualitative research, textual analysis, discourse analysis, and symbolic interactionism. Pre-service teachers are grouped jigsaw style into six expert groups. All groups view the same video of teaching performance, but each group focuses on one of six different codes: action/sequence, semes/forms, sound/discourse, proximity/movement, culture/context, and effects/process. Each group analyzes the video for their assigned code at three levels: literal observation, interpretation, and evaluation/application. When each group shares their focused analysis, the entire class of pre-service teacher arrives at a richer understanding of the teaching performance video.

Video cases are another type of video-based instructional approach in teacher education. Video cases are perhaps the most broad and widely researched type of video use in teacher education. I am using the term video case to represent what is actually a broad and diverse field of instructional approaches (L. Shulman, 1992). What these approaches have in common is the use of video to present a context-rich representation of teaching/learning interactions about which pre-service teachers can reflect.

Shulman (1992) discusses the multitude of variations on this theme. For example, there is the business school model in which the *A Case* presents the basic circumstances of the case along with a set of alternative courses of action that might be taken. After students are given an opportunity to reason about the *A Case*, they view the *B Case* which adds more information and typically includes an account of how the actors chose to actually respond

and the ensuing consequences. An alternative format for the video case is the narrative form in which the entire case is presented all at once as a process of discovery—complete with dead ends, insights, and uncertainties. Different video case formats vary in terms of amount of student-student and teacher-student dialogue, instructional purposes (illustration of principles, establishing precedents, establishing vision, suggesting strategy, reasoning about ethics, etc.), and extent of explicit structure for inquiry.

Chan and Harris (2005) have studied a unique class of video cases which they call *video ethnography*. Video ethnography integrates text with video "probes" or segments of video for in-depth inquiry. Video ethnography is distinguished by four design principles:

1. Authenticity (Cases involve an ethnographic approach rather than contrived, fictional case material.)
2. Problem-representation (Video ethnography allows the user to formulate the problem statement and solution paths.)
3. Presentation of multiple perspectives (Video ethnography includes various voices that comment on seen and unseen elements embedded in the context.)
4. Juxtaposition of theory and practice (The voice of research-based theory is layered into the video ethnography.)

Chan and Harris showed that video ethnographies scaffolded teachers' cognitive processes in ways that prompted development of teacher cognition.

Another way in which video has been used in teacher education is with multimedia essays, sometimes called video papers (Pea, 2002; Preston, 2005). A multimedia essay includes text with embedded links to "quoted" video clips. Video essays make a significant departure from video cases or ethnographies. For video papers, pre-service teachers are not consumers but rather principle authors of the video-based inquiries.

The Columbia Center for New Media Teaching and Learning developed a software tool called VITAL (Video Interactions for Teaching and Learning). VITAL includes a digital library of videos and scholarly articles, a video clipping/annotation tool, and a workspace for writing multimedia essays. In a preliminary study on the use of VITAL in the creation of video essays on the development of learner thinking in mathematics, researchers found that students used video clippings in three ways: as evidence, for analysis, and for connection (Preston, 2005).

Across the various uses of video in teacher education, a key purpose for including video is to improve the quality of dialogue and reflection around observed teaching events. One study considered the ways in which providing video essays, in addition to a more traditional text-only essay, influenced the discourse of teachers talking about teaching/learning events (Nemirovsky, 2005). When pairs of teachers were able to reference a multimedia version of a case study during discussion, their talk reflected two types of discourse: *grounded narrative* and *evaluative discourse*.

Grounded narrative has the aim of creating a narrative to describe and account for classroom events. Evaluative discourse focuses on the values, virtues, and commitments at stake in classroom events. An interesting feature of grounded narrative was the way in which participants used video to "zoom in and out" of the classroom events. The authors explained this zooming phenomenon this way:

> Zooming in entails dwelling in a particular event while keeping in mind the broader occasion, pulling apart the different aspects involved, and sometimes developing a micro-narrative that takes place 'within' the event. Zooming out establishes direct continuity across events separated in time while shifting utterances occurring in-between to the background. (Nemirovsky, 2005, p. 379)

Also unique to grounded narrative was the tendency for teachers to use the video to distinguish between what they had actually seen and what they had inferred. These findings indicate that providing video to teachers supports expert-like reasoning and perception.

Research has also begun to validate the value of video in teacher education in other ways. In one review of the literature on the relationship between video-technology and teacher education reform, video was found to support pre-service teachers in transforming their own beliefs, acquiring pedagogical content knowledge, and developing pedagogical understanding of learners (Wang, 2003). The use of video in teacher education for reflecting on others' teaching performance continues to show promise.

Feedback in Field Experiences

Although reflecting on others' teaching performance is valuable for helping pre-service teachers develop expert-like thinking skills, pre-service teachers still need the opportunity to practice improving their own teaching performance. Teacher education programs typically include some type of teaching practicum, often referred to as a field or clinical experience. Field experiences help pre-service teachers transform their theoretical or observation-based knowledge into performative knowledge.

To optimize the value of field experiences, teacher educators generally provide some type of feedback mechanism that helps pre-service teachers revise their teaching performance. Traditionally, this feedback has been provided by an external evaluator such as a university supervisor or a school-based cooperating teacher. This "expert" external evaluator typically observes the pre-service teacher's teaching performance and then both pre-service teacher and observer engage in debriefing following the lesson or teaching event.

During observation, the external evaluator often keeps some kind of record of significant teaching events. One traditional method that has been favored in physical

education teacher education is *systematic observation* (Everhart, 1996). In systematic observation, the external evaluator has a list of key teacher and/or student behaviors. While observing the pre-service teacher, the evaluator keeps a tally of the frequency of these behaviors. Following the performance, the evaluator can discuss the implications of the coded behaviors for future teaching.

With the transition from purely behavioral theories of learning to more cognitive ones, observation has also shifted focus to attend to the internal mental processes of students and teachers. This has led to the rise of observational methods that employ narrative or reflective methods. However, teacher educators have noted the weaknesses of these more purely qualitative approaches which tend to be unsystematic and overly subjective (Sharpe, 1997). Some practitioners have opted for a blended approach which combines both quantitative behavioral observations and qualitative narrative commentaries (Sharpe, 1997).

Even with these methods for providing external evaluation, teacher educators are not satisfied with the feedback that pre-service teachers receive while in the field. One significant problem is the ratio of supervising teachers to pre-service teachers severely limits the amount of feedback supervisors can give to each individual pre-service teacher (Struyk, 1993). Furthermore, the social differential between supervising teachers and pre-service teachers often creates a level of anxiety that can interfere with the pre-service teacher's ability to engage in developmental practice.

One solution to this problem has been to pair pre-service teachers for joint placement in a field experience. Researchers have established the social and learning benefits of partnered teaching in field experiences (Bullough, 2002). However, the peers of pre-service teachers may not have the necessary skills to provide accurate feedback and the result may be reinforced practice of ineffective teaching performance.

Filters in the Evaluation of Field Experiences

The problems of peer evaluation bring to the fore the fact that evaluation, whether conducted by a peer or by a supervising teacher, is a highly interpretive process (Nicol, & Macfarlane-Dick, 2006). Researchers have demonstrated that individuals draw upon both auditory and visual cues when evaluating teaching performance (Schmidt, 1994). Clearly evaluation, like other cognitive processes, involves the construction of a mental representation. This representation is actually the composite of the way an individual represents the performance and the way the individual represents the criteria by which that performance is to be evaluated.

Because evaluation involves representation, the quality of the evaluative feedback a pre-service teacher receives is influenced by a set of representational filters, such as attention and memory. Because teaching is a simultaneous, multidimensional, and immediate performance, pre-service teachers' awareness of their own performance is limited. Amidst the needs to consider proximate pedagogical moves, attend to classroom management, and assess student learning, teachers have limited processing capacity (Hogan, 2003; Sherin, 2004; Struyk, 1993). Information is inevitably lost.

This effect is particularly pronounced for novice teachers who have not yet developed automaticity for classroom procedures; however, even external evaluators are influenced by representational filters. External evaluators also have limits on their processing capacity and ultimately can only pay attention to a limited set of phenomena (Struyk, 1993). In a lively classroom setting, when an observer chooses to "zoom in" on a particular table of students engaged in group work, the observer has also selectively filtered out the experience of students sitting at another table. Given the typically unfavorable ratio of supervisors to pre-

service teachers, external evaluators' representations of performance are also subject to the effect of sampling.

After the performance, there are likely to be significant differences in the ways the external evaluator and the pre-service teacher represent the performance. Not only has the performance been filtered by attentional processes, but it is also now subject to the influence of memory (Sharpe, 1997). The effects of these filters help to explain the finding that pre-service teachers struggled to make their post-performance reflections specific. Not only were their reflections vague, but these teachers also gave unrealistic ratings of their own performance (Jensen, 1994).

The feedback a pre-service teacher receives from an external evaluator is also subject to a set of communicative filters, including language and generalization. As the evaluator attempts to talk about his/her own personal representation of the performance and evaluative criteria, the pre-service teacher actively interprets the evaluator's feedback in terms of his/her idiosyncratic mental representation of the performance and evaluative criteria. Here, language serves as a representation of a representation. For the novice teacher whose teaching vocabulary is still developing, language can introduce another layer of ambiguity (Stevens, 2003).

Generalization also plays into the communication of feedback. When debriefing pre-service teachers, external evaluators are likely to summarize their observations. In fact, Struyk (1993) has pointed out that most pre-service teacher evaluations in use today are designed to provide summarized results. The specificity of feedback that might improve teaching by focusing on specific teaching events is lost.

Given this daunting set of representational and communicative filters, it is remarkable that pre-service teachers do as well as they do with the feedback they receive. Nonetheless, there are obviously opportunities to improve the feedback process.

Video holds promise for reducing the effect of these filters. Goldman-Segall (1994) argues that video is a liminal medium that is simultaneously "raw" source data (content) and representation (artifact). Thus a video of a teaching performance is such a high fidelity representation of the performance that it can almost be considered the performance itself. At the same time, I recognize that video is itself an interpretation of an event as framed by a lens and point of view (Jones, 2004); however, the fidelity of this representation is still much more sensory-rich and accessible than the mental and verbal representations framed by pre-service teachers or external evaluators.

Because video, digital video in particular, changes the nature of a performance from something that is time-bound to something that is reviewable, the influence of representational and communicative filters can be greatly reduced (Sherin, 2004). Representations of the performance that have been filtered by limited attention during the actual performance can be revised and expanded by reviewing the performance. Inaccurate recall of the performance can be checked against the video record. Pre-service teachers and external evaluators can refer to a shared visual experience. Generalizations can be grounded in specific instances.

Self-Evaluation in Field Experiences

Even when external evaluators take advantage of the affordances of video for enhancing feedback, they are still limited by the amount of time they can spend with each pre-service teacher. One way to increase the amount of feedback that pre-service teachers receive is to train them to give themselves feedback. The idea of establishing a

culture of self-assessment is not original (Thomson, 1992). Teacher educators have explored various tools for supporting pre-service teachers in self-assessment (Poole, 2003). Ultimately, they recognize that pre-service teachers need to establish a self-regulatory practice that will continue to inform teaching performance long after students have become in-service teachers. Furthermore, researchers have established that pre-service teachers are willing learners who want to improve their teaching performance and will engage in self-evaluation if only they are given the necessary tools (Struyk, 1993).

Effective tools for supporting self-evaluation will take into account what is known about self-regulation. Zimmerman (2001) has pioneered research on self-regulation, identifying three key subprocesses: self-observation, self-judgment, and self-reaction. These are not mutually exclusive processes, but rather they involve a great deal of mutual interaction. Self-observation is of particular interest in the context of video feedback. Self-observation allows learners to assess their own behavior on dimensions of quality, quantity, and originality. The most effective self-observation involves self-recording that links instances of behavior to the contextual features in which they occur.

Nicol and Macfarlane-Dick (2006) synthesize research on self-regulation and formative feedback in higher education. They argue that students already engage, at least to some degree, in self-monitoring. To make this monitoring more systematic, teachers need to provide structured opportunities for self-monitoring. Nicol and Macfarlane-Dick (2006) describe seven characteristics of effective feedback:

1. It helps clarify what good performance is.
2. It facilitates the development of self-assessment in learning.
3. It delivers high quality information to students about their learning.
4. It encourages teacher and peer dialogue around their learning.

5. It encourages positive motivational beliefs and self-esteem.
6. It provides opportunities to close the gap between current and desired performance.
7. It provides information to teachers that can be used to help shape teaching.

Practices that support these aspects of feedback will not only improve learner practice but will also empower learners for ongoing development.

Video can help integrate many of the key characteristics of effective feedback into a self-evaluation process in teacher education field experiences. Video is particularly valuable as a tool to support self-recording and self-monitoring. The high quality representations of teaching performances afforded by video provide specific and contextualized information to pre-service teachers. Video provides an opportunity for pre-service teachers to revise their internal representations of their own performance, thus more accurately identifying their own performance gap.

Video Support for Self-Evaluation

The tension between structured quantitative observation and flexible qualitative reflection that is evident in the historical evolution of pre-service teacher field supervision (Sharpe, 1997) is also evident in the spectrum of video uses for self-evaluation. Video evaluation strategies have tended to evolve from the narrowly defined, systematic counting of classroom behaviors to the broad, subjective exploration of learning through reflection. Ultimately, developments in video technology are pushing towards a blended approach.

One of the earliest and most widely researched applications of video to teacher education was the practice called microteaching (Sherin, 2004). Microteaching was an attempt to provide practice with a simplified teaching task to pre-service teachers. During microteaching, the pre-service teacher would give a short lesson of about five to ten minutes in length that focuses on a single teaching skill. Following the practice, the pre-service

teacher reviewed a videotape of his or her lesson to analyze his or her use of the teaching skill. Afterwards, the pre-service teacher might revise the lesson and present it to a new group of students, again followed by video feedback. For microteaching, self-evaluation was highly systematic, not because the video and analysis process was necessarily well-defined but because the teaching experience was tightly constrained.

The influence of microteaching can still be seen in teacher education field experiences. In a more recent study, pre-service physical education teachers engaged in one-on-one lessons with the same learner twice a week for a series of six weeks (Brawdy, 1994). Following each lesson, the pre-service teachers analyzed a video recording of their performance, using systematic observation to assess the way they had used feedback to support student learning. Thus much like micro-teaching, the domain of these teaching experiences were highly constrained. Furthermore, the introduction of systematic observation into the lesson introduced another level of systematic information gathering for feedback purposes. Because the study compared self-analysis to analysis by a supervisor, the researchers were able to determine that the ideal model of evaluation involved a preliminary period of independent self-assessment.

In another study, pre-service teachers used both systematic observation techniques and systematic filming techniques (Struyk, 1993). The pre-service teachers viewed video of their own teaching in which the camera was panned across the classroom in twenty second intervals. The pre-service teachers analyzed the video by coding and counting the behaviors that were observed during this systematic video sampling. The authors of this study argue that this method was successful for providing feedback to pre-service teachers.

In contrast to these highly constrained uses of video for evaluation, other teacher educators have used video as a text for more flexible interpretation and reflection. In one

study, pre-service teachers learning to teach about writing each produced a film of themselves describing and teaching a specific writing strategy (Collins, 2004-2005). Videos were collaboratively reflected on by peers in a loosely structured process called *critical friends*. All student videos were finally compiled to a CD that could serve as an experience bank for course participants.

However, even video reflection can be improved by introducing at least some degree of structure. Pre-service teachers in one study videotaped three of their own teaching sessions (Jensen, 1994). Using these videos, the teachers engaged in focused reflection on one of three areas for each video: interpersonal skills, instructional management and organizational skills, and questioning skills. Following all three reflections, they completed an assessment of their overall teaching competencies. Study results suggested that pre-service teachers' focused reflections on their own teaching provided more reliable and useful information than their assessments of their overall teaching competency.

Sherin and van Es (2005) have taken focused reflection to a new level by introducing a video analysis software that supports reflection processes. Their tool, called Video Analysis Support Tool (VAST), scaffolds video viewing by prompting analysis in three teaching domains: student thinking, teacher's roles, and classroom discourse. For each of these domains, users are asked what they noticed, what evidence they had found, how they interpreted the events, and what other questions were raised. Pre-service teachers who used VAST to analyze videos were found to shift focus from describing the chronological range of all events to identifying the significant/relevant features. They also experienced shifts from an evaluative to an interpretive mode and towards a more evidence-based approach to reflection.

Reflective processes have also been scaffolded by involving pre-service teachers in the editing of videos of their own teaching performance. As discussed previously, one approach to reflection on others' teaching is to provide a video bank from which teachers select video clips to construct a multimedia essay. In the same way, video of pre-service teachers' own teaching is also being used as material for case studies. Nicol and Crespo (2004) describe their work with a course on problem-based education. Pre-service teachers in the course were assigned to produce a case study of teaching and learning based on their practicum experience. The authors described the experience of a single, representative student from their course, focusing on what she attended to in the process of filming/editing and describing/analyzing her video. Findings indicated that she moved from wondering what to film to attending to students' mathematical dispositions and then to seeing parallels between her own learning and that of her students.

In a related pilot study, pre-service teachers edited video of their own teaching in order to communicate their growth as teachers (Cunningham, 2002). Although these teachers reported that the editing process encouraged them to engage more attentively and critically with their videos, they also found the editing process to be time consuming. The time intensive process of video editing reduced the amount of time students could spend in meaningful reflection.

Video Annotation Tools

This literature suggests several ways in which video feedback can be enhanced. First, teacher educators need to focus pre-service teachers' attention on key aspects of teaching by explicitly structuring evaluation/reflection activities. Second, pre-service teachers need the opportunity to anchor their observations by referencing specific events in their videos. Third,

teacher educators need a time-efficient method for helping students meet the first two requirements.

An emerging family of video annotation software holds promise for supplying these conditions. Video annotation circumvents the problems associated with teaching learners to edit videos. With this type of software, teachers can define a set of annotation codes or tags that focus student learning on key issues. Learners can then use the intuitive annotation tools to simply tag or markup a video. It is almost as simple as using a highlighter or red pen to make comments on a term paper. Video annotation software may provide a practical solution to facilitating effective use of video for self-evaluation in teacher education.

Before I discuss how video analysis tools might be used in teacher education, I want to point out that video analysis tools are emerging in a number of different fields and disciplines. For example, a video annotation tool called MediaTagger has been developed in the field of psycholinguistics ("Max Planck Institute for Psycholinguisitics: General description Media Tagger"). This tool allows the user to select a time slice of video and tag it with a transcription or code. These tags can be organized onto different layers for different analytic purposes. A similar tool called ANVIL has been developed in Germany (Kipp, 2001). Although this latter tool is intended to be a generic video annotation tool, the developers are using the tool to support creation of agent interfaces (i.e., animated characters).

In museum education, Stevens (2001) has led the development of another video analysis tool called VideoTraces. This software is unique in that it allows the user to annotate video with verbal comments, by adding a secondary audio track to the video, and gestures, by recording the movements of a mouse pointer tool on top of the video display. This tool also allows the user to alter the playback speed while creating these annotations, or

video traces as Stevens calls them. Stevens has piloted the use of VideoTraces in science museums. By allowing museum visitors to create video traces about their experiences with and thinking about particular exhibits, he hopes to facilitate a dialogic and constructivist learning environment.

VideoTraces has also been used in an undergraduate course on dance choreography (Cherry, 2003). Because dance is a performance-based discipline, the instructional problems associated with teaching dance choreography are similar to those faced by teacher educators. The researchers who piloted the use of VideoTraces for teaching dance explain the problems the software is intended to address:

> When communicating about a dance, students and instructors must remember the details of the dance, and must coordinate their comments with the actual elements of the piece, without a concrete representation to which to refer. Without a common external point of reference, conversation relies on the possibly faulty assumption that the participants in the conversation have a shared mental model of the dance. (Cherry, 2003, p. 1)

In other words, dance teachers, like teacher educators, also deal with representational and communicative filters when providing feedback to learners. These researchers make preliminary claims about the effectiveness of using a video annotation tool like Video Traces for teaching performance arts.

Little research has been conducted on the effectiveness of using video annotation tools in teacher education. However, research on other video analysis tools and methods suggests that video annotation may be a powerful way to enhance pre-service teacher learning. At BYU, Physical Education teacher educators are piloting the use of a video annotation tool called Studiocode to help pre-service teachers self-evaluate their teaching

performance. Anecdotal evidence suggests this is an effective and efficient way to provide pre-service teachers with useful feedback about their teaching. However, researchers need to systematically investigate these claims.

Other researchers have argued that teacher education research needs to compare methods of self-assessment with traditional methods of external assessment (Jensen, 1994). Before researchers can investigate the relative value of video annotation as a method for improving the actual teaching performance of pre-service teachers, we need to first determine to what extent students can reliably apply annotation codes to videos. Reliability issues are one of the key factors that limit the implementation of both student- and teacher-performance assessments (Bitner, 1995; Pecheone, 2005). If pre-service teachers do not provide themselves with reliable feedback information, then it would be difficult to validate claims that video annotation improves teaching performance.

One pilot study investigated the ability of induction teachers to accurately use a video analysis tool to identify examples of quality teaching in videos of their own teaching (West, 2007). This study reported that most of the induction teachers who participated were able to identify examples of each of the targeted attributes of teaching quality. Furthermore, the study found that teachers were better able to identify examples of the teaching attributes after participating in group reflection and discussion about specific video examples. Although this study provides encouraging preliminary support for the ability to train teachers to be reliable video analyzers, the study is limited by its method. It relies mainly on focus group data which is subject to self-report errors. Furthermore, the study does not systematically and quantitatively investigate the reliability of teacher coding. In the present study, I will expand upon the research conducted in this preliminary study.

Chapter 3: Method

Participants

Participants for this study were recruited from two sections of the course Exercise Science 375, Physical Education for Elementary School Teachers (EXSC 375), at Brigham Young University. Of the 54 students enrolled in EXSC 375 during the first block of the Winter 2007 semester, 49 students agreed to participate in the study. From Section A, 23 students participated and 26 students participated from Section B. All participants signed informed consent forms agreeing to participate. Of the participants, 48 were female and only 1 was male.

Instructional Context and Intervention

Two instructors teach different sections of EXSC 375. The course is a 2 credit hour block class, meaning that for eight weeks, the class meets twice a week for two-hours each session. This schedule allows students the time they need to go into local elementary schools to teach PE lessons. They do this a total of four times during the course. The first three times they team-teach the PE lesson with a partner who is also enrolled in the class. The final time they solo teach a PE lesson. Although this schedule makes it possible for the students to have valuable field teaching experience, the accelerated nature of a block class, which is only half the length of a traditional semester, presents challenges for instructors who must cover a lot of curriculum very quickly. For elementary education majors, this is one of the first education classes these students take in the course of their program's curriculum. Thus one of the purposes of the class is to provide students with a rapid and intense introduction to classroom management.

An integral part of the course is the use of a video analysis software tool to analyze videos of the students' teaching in the field. For each of the four PE lessons, students capture

their teaching on digital video. Afterwards, they use a video analysis software tool called Studiocode to analyze their teaching (see Figure 4). This software provides a set of instructor-defined codes which can be used to index events in a video. When a code is attached to an event, the user has the option of also attaching written comments.

For EXSC 375, the instructors have developed a set of codes that represent desired teaching competencies that students should develop by the end of the course. These desired competencies have been derived from a careful consideration of some of the best practices of expert PE instructors. This list of competencies is not intended to be comprehensive, nor do the instructors pretend that mastery of these competencies will ensure teaching excellence. Nonetheless, the competencies described by this set of codes constitute a set of skills instructors believe will empower students in their class to have at least some success in teaching PE as they continue to develop greater, more flexible levels of expertise. Because of the activity-based nature of PE teaching and the timing of EXSC 375 at the beginning of the elementary education program curriculum, these codes focus on competencies related to classroom management.

The set of codes defined for EXSC 375 include five main codes with fifteen sub-labels (see Table 1). A *code* is an overarching category that is used to tag a segment of video. Each code may include several labels which classify or qualify the code. For example, the first code is "freeze positions." This code is used when the PE teacher gives a signal, such as the blow of a whistle, and the children are expected to stop whatever they are doing, turn towards the teacher with their hands on their knees and without talking await further instructions. Under the "freeze positions" code there are two labels which are used to describe how children respond to the teacher's freeze signal. If all students in the class respond appropriately, then the "Full compliance" label is used. If some of the students do

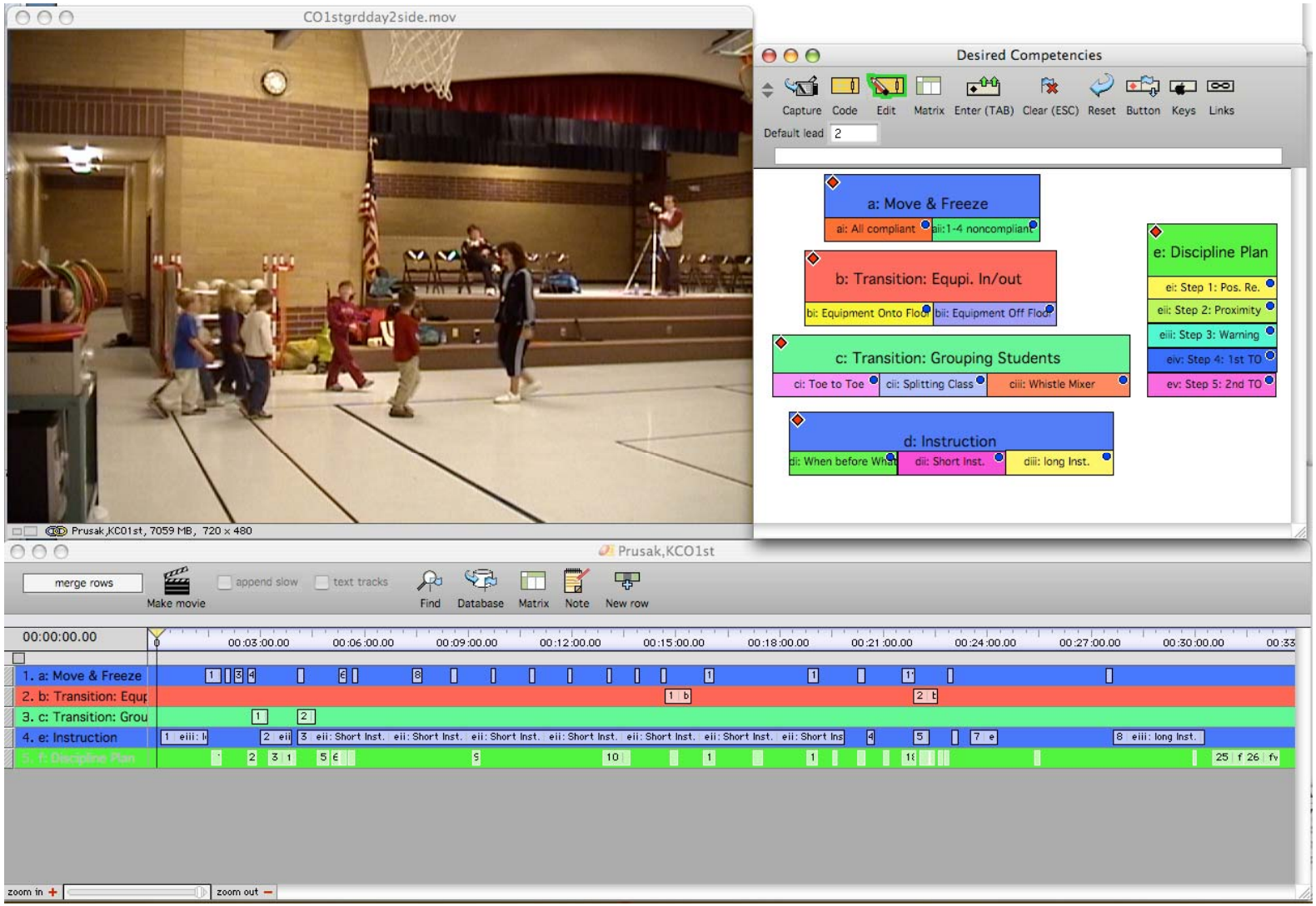

Figure 4. Screenshot of Studiocode with the video viewing window (upper left), coding buttons (upper right), and coding timeline (bottom).

Table 1

Description of ExSc 375 Coding Scheme

Code	Label	Component Parts
a. Freeze Positions	i. Freeze - All Compliant	1. Teacher on the perimeter 2. Model the freeze 3. Quietly scan side to side 4. Use positive praise; specific to all parts of the freeze position to more than one student 5. Stand them tall right before giving the next instruction
a. Freeze Positions	ii. Freeze – 1-4 Non-compliant	1. All of the above components (freeze-all compliant) 2. 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. Equipment is set up spaced around the perimeter of the floor 2. Use When before What 3. Model how to get the equipment and find your own space. Then model the first activity they should do (transitional activity) 4. Completed in a timely manner (after the go, no more than 10-15 seconds until all students are engaged)
b. Transitions involving Equipment	ii. Moving equipment off the floor	1. Reverse the above procedure 2. Emphasize and model carrying the equipment (not playing with it) and placing it down (not throwing or rolling it) 3. Give students a transitional activity for after the equipment is put away 4. Complete in a timely manner (after the go, no more than 10-15 seconds until all students are engaged)
c. Transitions moving students	i. Making partners (toe to toe)	1. Freeze the class first, scan, and reinforce 2. If teaching the first time 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" 3. If they know Toe to Toe already just say the words "Toe to Toe" 4. After the "Go", remind those that don't have partners to hustle to the middle, raise their hand, and find a partner 5. Completed in a timely manner (10-15 seconds after the go)
c. Transitions moving students	ii. Splitting the class into two teams	1. Use toe-to-toe as described above 2. Have one person sit down 3. Model where they should go and how they should stand, then move the standers first while leaving the sitters in place 4. Once the standers are in place, move the sitters to where they should go 5. Completed in a timely manner (under 30 seconds)
c. Transitions moving students	iii. Making groups (whistle mixer)	1. Freeze the class first, scan, and reinforce 2. If teaching the first time 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." 3. If they know Whistle Mixer then just blow the whistle the appropriate number of times. 4. After the Go, remind those that don't have a full group to hustle to the middle, raise their hand, and find a full group 5. Completed in a timely manner (10-15 seconds after the go)
d. Instructional episodes	i. Using "When" before "What"	1. Use the phrase "When I say Go, I would like you to . . . , Go" (do not use the whistle to start) 2. Model the activity you want them to do during your explanation and before you say, "Go" 3. 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, . . ."
d. Instructional episodes	ii. Giving short instructions (< 1 minute)	1. Freeze the class, scan, reinforce and **stand them tall** 2. Give the instructions using concise teacher talk, direct and to the point. Do not compete with talking students. Model the instructions 3. When appropriate use the phrase "When I say Go, . . ." to get the class started again 4. After the "Go" scan, check cards if necessary, and reengage with the class working the crowd

Table 1 (continued)

Description of ExSc 375 Coding Scheme

Code	Label	Component Parts
d. Instructional episodes	iii. Giving long instructions with students sitting down (> 1 minute)	1. Freeze the class, scan, reinforce and **sit them down** where they are 2. Give the instructions using concise teacher talk, direct and to the point. Do not compete with talking students. Model the instructions 3. When appropriate use the phrase "When I say Go, . . ." to get the class started again 4. After the "Go" scan, check cards if necessary, and reengage with the class working the crowd
e. Using the Discipline Plan	i. Step 1: Positive reinforcement	1. Use positive reinforcement to change non-compliant behavior 2. Clearly identify to the class a compliant student, use their name, and positively reinforce the specific behavior of the non-compliant student
e. Using the Discipline Plan	ii. Step 2: Teacher proximity	1. Continue teaching/reinforcing while you move next to the problem student(s) 2. Make note if the behavior is corrected, if needed, be ready to give a warning
e. Using the Discipline Plan	iii. Step 3: Quiet warning	1. Reengage the class 2. 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 rule number 1 and 2, this is your warning, next time you'll go to time out." 3. Walk away and reengage with the class
e. Using the Discipline Plan	iv. Step 4: First Time out	1. Reengage the class 2. 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." 3. Walk away and reengage with the class and give some positives. Take note if the student went to time out
e. Using the Discipline Plan	v. Step 5: Second Time out	1. Reengage the class 2. 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." 3. Walk away and reengage with the class and give some positives. Take note if the student went to time out.

not respond to the freeze signal correctly, then the label "1-4 noncompliant students" is used. Depending on which label is used, the PE teacher is supposed to then either praise the children's compliance or, for non-compliance, implement the discipline plan. Each code has several component parts (Hereafter "code" will be used to signify each of the 15 code labels. "Label" will only be used when the distinction between code and label is important.) In order for the desired teaching competency to be performed correctly, each of the component parts should be present. For example, to correctly implement the freeze, the teacher should be on the perimeter of the class when giving the signal to freeze, model the freeze position,

scan the children's behavior to check for compliance and to give the children an opportunity to bring their behavior into compliance, and then either praise their compliant behavior or implement the discipline plan. Finally, the freeze is completed when the teacher tells the students to "stand tall" before giving the next instruction. If one of these component parts is missing (e.g., if the teacher were to omit the scan) then the freeze would not have been implemented correctly. Thus this coding scheme is intended to attain a high level of operational definition that can be unambiguously understood by novice PE teachers.

Students in EXSC 375 first practice using Studiocode to code videos of an expert PE teacher. After this training, students then use Studiocode to code videos of each of their own four teaching performances. For any coding session, Studiocode produces a matrix which shows how many instances of each code were marked in the video. Students in EXSC 375 turn in a copy of this matrix to their instructor for each video they code. For this portion of their assignment, grades are based on completion of the assignment, not on the quality of coding. At the end of the semester, students in the class select the best example of each desired competency from their own teaching videos. They compile their best examples into a single video timeline that they turn in to the instructor. The instructor grades students based on the quality of teaching in these "best instances." Thus the primary objective for using Studiocode in the class is to focus on actual teaching competency, the ultimate goal of teacher education.

Data Collection and Analysis

This study included two phases. First, students participated in a coding proficiency test in which they used Studiocode to analyze a video of an expert teacher leading a PE lesson. Second, researchers conducted an audit of the students' coding of videos of their own teaching.

Coding proficiency test. During the first phase of data collection, researchers conducted a coding proficiency test that compared the coding abilities of student teachers and expert PE teachers. The two EXSC 375 instructors first selected a video of an expert teacher giving a PE lesson. This video was selected because in it the expert teacher demonstrated a majority of the codes (13 of the 15 possible codes). Each of the instructors, acting as expert PE teachers, then independently coded the video. Following independent coding of the video, the instructors compared the matrices from their coding. This allowed them to compare the number of each code used in coding the video. They collaboratively 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 of expert coding for the video.

During the first several weeks of class, students in EXSC 375 received intensive training on the teaching competencies represented by the coding scheme. Not only did students view videos of an expert teacher demonstrating these competencies, but they also participated in live modeling of those competencies. During these live modeling sessions, the class met on a gym floor with actual PE equipment like that used in elementary PE lessons. The students played the role of children in a PE class while the instructor served as the PE teacher modeling each competency.

Measures were taken to ensure that both instructors gave uniform instruction about the desired teaching competencies. The two instructors collaboratively planned their instruction prior to the semester. These instructors had worked together teaching this class for several semesters prior to this study. Thus most differences between their course content had already been eliminated. The teachers resolved any residual

discrepancies by collaboratively defining the component parts of each code and formalizing these in a handout for students. During the course, the instructors engaged in frequent dialogue to ensure that they were presenting consistent information in their respective classes.

In addition to these measures, I attended both sections of the course to observe how each instructor presented the information about the desired teaching competencies. While attending the classes, I took field notes to record any potential differences between how the instructors taught the course. During these observations, I was particularly attuned to any differences in what the instructors did or said or activities the students did or did not participate in that might influence the way the students understood the proficiency test. That meant that I was attuned to how the instructors taught about each of the teaching competencies and how they introduced the Studiocode software. I was also alert to how the instructors positioned the proficiency test in terms of the purpose of the test and what it would require of the students. After completing the observations, I compiled my field notes and compared the instruction given by each instructor for each day of class. If an instructor covered a particular topic not covered by the other instructor on a particular day, I checked to see that the other instructor did or did not cover the same topic on another day. I summarized my observations by listing the significant similarities and differences between the two classes.

In addition to receiving instruction about the teaching competencies, students were trained in the use of the Studiocode software. They learned the basics of navigating a video timeline and applying codes to the video timeline. This training focused on the mechanics of using Studiocode (i.e., adjusting the start and stop times of a code or

changing the label affixed to a code), but it did not focus on the use of the specific coding scheme created for this course.

Following instruction and training, students participated in a coding proficiency test. Students were asked to code the video of the expert teacher giving a PE lesson—the same video that had been pre-coded by the instructors. Students were informed that their coding would be compared to the coding done previously by an expert coder. They were told to try to identify all instances of each code in the video. Their score would be deducted if they missed an event coded by the expert coder, if they mislabeled an event, or if they coded an event that was not coded by the expert. They were informed that their grade for this assignment was based on completion of the test.

Because students in Instructor A's class completed the proficiency test prior to the students in Instructor B's class, adjustments to the proficiency test protocol were made for Instructor B's class. Class A was told that they needed to reach 80% agreement with the expert coder. If they did not attain this level of agreement on their first attempt of the proficiency test, students would be given feedback on their coding and could correct their coding and rescore their test. They could attempt the test as many times as were needed to reach the 80% agreement level.

Course instructors anticipated that students would be able to reach the 80% agreement level of coding competency within three attempts. However, when students in Instructor A's class struggled to demonstrate this level of coding competency after multiple attempts at the proficiency test, the instructors and researchers agreed to change the testing protocol. The majority of students in class A had already attempted the proficiency test at least three times and were demonstrating a high degree of collective frustration with the test. Thus instructor A informed his students that they did not need to

keep taking the test; they would receive full credit for the assignment. When Class B began the proficiency test, students were told that if they did not score 80% agreement on their first attempt, they should repeat the test up to three times; however, if by the third attempt they still had not reached the 80% agreement they were done with the test and would receive full credit for the assignment.

Scoring of the proficiency test was automated with spreadsheets so that students could score their own test and receive automatic feedback on their coding. To score the proficiency test, students exported their coding to a spreadsheet which listed each code's start and stop time. They pasted this data into a spreadsheet which compared each of their codes against the codes from the expert coder. This automatically generated two scores. The first score was the percentage of match with how the expert labeled the codes. The second score was the percentage of how much time overlap the students' coded video segment had with the expert's coded video segment.

The first score was calculated by taking the total number of codes used by the expert and deducting for deviations (i.e., extra and missing codes), then dividing by the total number of expert codes (see Figure 5).

$$\text{Percentage of Matching Codes} = \frac{\text{Total Number of Expert Codes} - \text{Number of Missing and Extra Codes}}{\text{Total Number of Expert Codes}}$$

Figure 5. Calculation of the matching code labels percentage.

Hereafter, I will refer to this score as the matching code labels percentage. Although this score told whether the student had coded the same event as the expert, it did not show how much of the same event the student was actually seeing. The matching code labels

percentage simply indicates that at some point the student's coded video segment overlaps with the expert's coded video segment, but it does not describe how much they overlap. This explains the need for a second score, which I will call the time overlap percentage. This score was calculated by dividing the amount of time that the student's coded video segment overlapped with the expert's coded video segment by the total length of the expert's segment (see Figure 6).

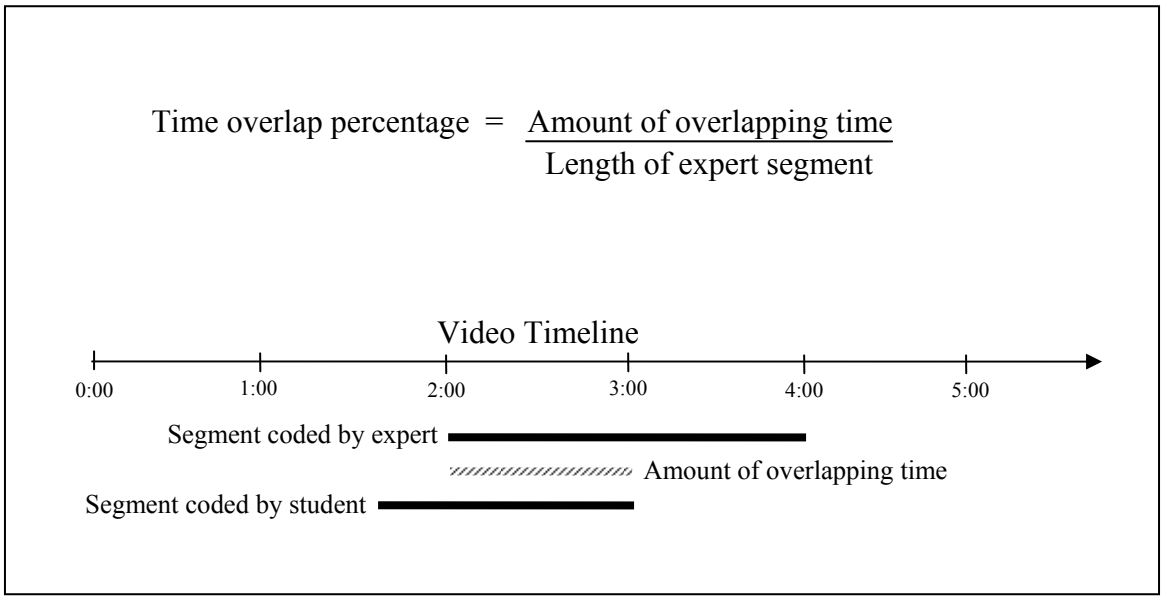

Figure 6. Calculation of time overlap percentage.

When calculating whether or not a student had demonstrated sufficient coding competency for the proficiency test, the spreadsheet required that both the matching code labels score and the overlapping time score reach at least 80%.

If the student did not reach this level of coding competency, the spreadsheet displayed feedback to the student. This feedback was displayed in two lists. The first list showed how many instances of each of the 15 codes the student was missing when compared to the expert coder. The second list showed how many extra instances of each code the student had used that did not match the expert. Using this information, students

could go back to their coding file in Studiocode and revise their coding of the video. Once revised, students could again attempt to score their proficiency test.

For each attempted scoring, students were instructed to save a copy of their scoring results. They identified each attempt as "Attempt 1", "Attempt 2", "Attempt 3", etc. so that researchers would later be able to determine their initial and final levels of coding performance. All Studiocode files were saved to a university computer server so that researchers would be able to access the data.

Students' scores from the proficiency test were analyzed using simple descriptive statistics. I calculated the mean score for both the matching code labels score and the overlapping time score for each of the first three proficiency test attempts. Because some students did not save each of their attempts to the computer server, there were some holes in the data set. Some students had only saved one or two of the three attempts. This explains why the *n*'s vary for each attempt. Despite this inconsistency in data collection, there are fairly consistent *n*'s for the three attempts (see Table 2). I decided to calculate the means based on any usable data that had been collected, regardless of whether the student had not recorded data for all three attempts. I decided to do this in order to avoid biasing the analysis towards those students who recorded all three attempts and might have been a little more attentive to details or classroom instructions.

As might be anticipated, the mean scores for each successive attempt showed improvement. Because the main variable between successive attempts was the influence of feedback and coding practice, I wanted to determine whether the perceived improvement might be attributable to growth or merely the result of random variance. To determine whether the difference between the scores for each attempt was statistically significant, I conducted a within-group, repeated measures ANOVA. This statistical

analysis allowed me to determine whether the observed improvement over the three proficiency test attempts was due to random chance or whether it was likely due to the systematic influence of practice. The logic of a repeated measures study is that the same participants are subjected to the same measure on different occasions. Because each of the repeated measures is for the same participants, the participants serve as their own control group. If a difference is observed between the repeated measures, it can be attributed to whatever treatment was given during the intervening period between measures. In order to conduct the ANOVA in accordance with this logic, it was necessary that I use only the data for those students who recorded all three of their proficiency test attempts. This reduced the n slightly to 27, but the *n* is only slightly different from the 30-33 participants for whom I had any recorded data. Thus this sample of the data can still be considered reasonably representative.

Table 2

Number of Participants Who Recorded Data for Each Attempt

	n (Attempt 1)	*n* (Attempt 2)	*n* (Attempt 3)
Both Classes	30	33	33
Class A	9	11	12
Class B	21	22	21

Because class membership was a potentially confounding variable in the data, I also calculated the mean scores for each proficiency test attempt by class. When I did this, I found that the mean scores for Class A appeared to be higher than those for Class B. To determine whether this difference was statistically significant, I conducted a between groups t-test. This statistical analysis allowed me to determine whether the observed difference between scores for the two classes was due to random chance or whether it was likely that the difference was due to class membership.

I also generated descriptive statistics to identify those specific codes that might have been particularly problematic for students. I did this by calculating the mean number of missing and extra instances for each of the codes. Because some of the codes represented teaching competencies that naturally occur with greater frequency in most PE lessons, I weighted these results by calculating the average amount of error per code. I did this by dividing the mean number of missing and extra instances of each code by the total number of instances of each code as identified by the expert coders. This analysis allowed me to determine which codes were most problematic in terms of having either too many (extra) or too few (missing) instances identified by students. By looking at these values, I decided to classify those scores at or above .40 as problematic. A score of .40 meant that for every ten instances of a particular code, there were four extra or missing instances of that code. This cut off point included approximately the highest third of the problematic codes and was a natural breaking point in the distribution of scores.

After both classes had completed the proficiency test, I conducted informal follow-up interviews of students and EXSC 375 instructors to help contextualize the proficiency test results. To select students to interview, I divided each class into thirds based on their performance on the proficiency test. I then invited students from the lowest and highest thirds to participate in the interview. A total of eight students were interviewed (4 from the highest and 4 from the lowest thirds, 3 from class A and 5 from class B). This purposeful sampling ensured that interview data would reflect a range of student experiences with the proficiency test. During these interviews, I asked the students and instructors to explain why students might have coded the video of the expert differently than the expert coder.

Following the interviews, I analyzed my interview notes using simple qualitative analysis methods. The purpose of the interviews was to explain why student reliability scores were so much lower than anticipated. Thus, I categorized student and instructor comments based on the type of explanation given. In other words, I categorized comments according to the type of coding issue or problem it described. As I conducted the interviews, I became aware that many of the explanations given by students to explain the variances in their coding had little to do with how well the students understood the teaching competencies and had more to do with technical aspects of how they implemented the coding system. Thus, I classified each explanation given by students and instructors based on whether the reason given had to do with actual understanding of the teaching competencies or technical aspects of implementing the coding system.

Self-coding audit. The second phase of data collection involved gathering information about how well students could code videos of their own teaching. For this phase of data collection, students captured digital video recordings of themselves leading PE lessons in local elementary schools. The videos used in this phase of data collection came from the fourth and final PE lesson given during this course. Readers may recall that during the first three lessons students team-taught in pairs, but during the final lesson each student solo taught a PE lesson. While one student taught, the partner operated the video camera. Students were not given uniform directions about what to capture on the video. However, most followed the pattern modeled by the videos of the expert PE teacher in which the camera followed the teacher around the gym floor. Thus the student teachers were almost always on camera. This also meant that some of the children in the PE classes were typically off camera.

Students used Studiocode to code their own teaching videos with the same 15 codes used in the proficiency test. Unlike the proficiency test, students were not instructed to exhaustively code all instances of each teaching competency. Thus the same method of scoring the coding was not possible. Instead, researchers conducted an audit of the codes to determine whether the codes students used were correct identifications of the coded teaching competency.

For the audit, researchers selected a random sample of twenty student videos (10 students from each section of the course). Due to a high level of technical difficulties in camera operation or video storage, a large number of student videos were deemed unusable. One of the most common problems with the videos was that the audio track was out of sync with the video track. This problem was probably due to the way videos were compressed by the Studiocode software. Another common problem was the complete lack of any audio recording. This was probably due to user error in camera or microphone operation. In total, 29 student videos were classified as usable. Of those 29 videos, 20 were randomly selected for the audit.

For each video, a researcher randomly selected 15 coded instances. For each instance, the researcher determined whether or not the code label was an accurate representation of the video segment. If the code was used appropriately, the researcher scored it as "correct." If the code was misapplied, the researcher scored it as "incorrect" and made note of why the code was inaccurate. This method for scoring only allowed the researcher to check for the accuracy of positive identifications of coded instances. Because the students did not exhaustively code the videos, the researcher could not determine whether or not a student could discern every instance of a teaching competency.

Chapter 4: Findings

This section provides an analysis of the findings on the reliability of coding with Studiocode in EXSC 375. It begins by considering the findings on the reliability of the two expert coders (i.e., EXSC instructors) with each other. This will demonstrate whether an expert standard of coding can be reliably established as a criterion for evaluation of novice coding. Then, the reliability of novice coding of an expert teaching video will be considered in light of findings from the coding proficiency test. Results from the self-coding audit will shed further light on the coding ability of novice teachers, particularly as it relates to being able to reliably code videos of their own teaching performances. Finally, the results from both the coding proficiency test and the self-coding audit will be analyzed to explore the potential sources of variance in coding reliability scores.

Reliability of Expert Coding

Prior to administering the coding proficiency test to students, the two instructors for EXSC 375 each independently coded the video of the expert PE teacher. Following this independent coding, the instructors collaboratively compared their coding and through dialogue resolved discrepancies in how they coded the expert video. Through this process, the instructors arrived at a single, agreed upon Studiocode file that represented an expert coding of the video.

Unfortunately, Instructor A's initial Studiocode file was not saved. However, Instructor B's Studiocode file was retained and can serve to illustrate the degree of agreement between the instructors. Instructor B's individual coding file was scored against the joint instructor Studiocode file. The percentage of matching code labels for Instructor B's individual coding is 93%, and the percentage of overlapping time is 96%. These high scores demonstrate a high degree of inter-rater reliability.

If Instructor A's initial Studiocode file had not been lost and the true inter-coder reliability could be calculated, it is likely that the initial reliability would be slightly lower. These reliability percentages reflect comparison of an independent coding with a collaborative coding. There would likely be more variance in the comparison of two independent codings. Nonetheless, the amount of variance between Instructor B's coding and the joint coding is so small that even if it were doubled, one way to estimate the variability between two independent scores, the percentages would still be 87% (matching code labels) and 92% (overlapping time)—well within the range of acceptable inter-coder reliability. Thus the high level of reliability between the two expert coders provide a basis for using the expert coding as a standard for evaluating novice coding reliability.

It will be helpful in interpreting the results of the coding proficiency test to identify the sources of variance in the way the two instructors coded the video. The instructors' codings were divergent on four codes. Comparing Instructor A's individual coding with the joint coding file shows a mismatch on three of the missing codes (see Table 3).

Table 3

Instructor B's Coding Compared with Joint Instructor Coding

Code	Extra	Missing
ai: Freeze - All compliant	0	0
aii:Freeze - 1-4 noncompliant	1	2
bi: Transition - Equipment Onto Floor	0	0
bii: Transition - Equipment Off Floor	0	0
ci: Student Transition - Toe to Toe	0	0
di: Instruction - When before What	0	0
dii: Instruction - Short Instruction	0	0
diii: Instruction - long Instruction	0	0
ei: Discipline - Step 1: Positive Reinforcement	0	0
eii: Discipline - Step 2: Proximity	0	0
eiii: Discipline - Step 3: Quiet warning	1	0
eiv: Discipline - Step 4: 1st Timeout	0	0
ev: Discipline - Step 5: 2nd Timeout	1	0

Instructor A had one extra instance of each of the following codes: *ai. Freeze – 1-4 non-compliant*, e*iii. Discipline Plan – Step 3: Quiet Warning*, and e*v. Discipline Plan – Step 5: Second Timeout*. Furthermore, Instructor A had two missing instances of the code *ai. Freeze – 1-4 non-compliant*. In addition to these disagreements between how the two instructors coded the video, interviews with the instructors revealed that initially their coding had disagreed on two other codes. One of the instructors had failed to code five instances of d*ii. Instruction – Short Instruction* and one instance of *ev. Discipline Plan – Second Timeout*. In summary, the instructors showed some variability in the way they used the following four codes:

1. ai. Freeze – 1-4 non-compliant
2. dii. Instruction – Short Instruction
3. eiii. Discipline Plan – Step 3: Quiet Warning
4. ev. Discipline Plan – Step 5: Second Timeout

Coding Proficiency Test Results

The results from the proficiency test show three things: first, the lower than expected reliability of novice coding; second, the improvement of student scores with practice; and third, the impact of the instructor variable.

First, students were not as reliable coders of the expert video as researchers and instructors had anticipated prior to this study. Compared to the high reliability between expert coders, novice coders demonstrated moderate reliability in their coding of the expert video in the proficiency test (see Table 4). On the first attempt, the mean percentage of matching code labels was 49%. The mean percentage of overlapping time was about five percent better at 55%. By the third attempt, these percentages had improved to 63% (matching code labels) and 68% (overlapping time). Despite the

improvement over the three attempts at the proficiency test, the scores did not approach the expected 80% reliability.

Table 4

Coding Proficiency Test Results

		Attempt 1			Attempt 2			Attempt 3		
		n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
All participants	Matching Code	30	49%	17.70	33	57%	18.32	33	63%	20.88
	Overlapping Time	30	55%	12.50	33	65%	10.34	33	68%	10.85
Class A	Matching Code	9	63%	16.15	11	62%	19.32	12	76%	20.80
	Overlapping Time	9	63%	8.31	11	65%	9.84	12	73%	13.98
Class B	Matching Code	21	43%	15.23	22	55%	17.77	21	56%	17.43
	Overlapping Time	21	51%	12.29	22	64%	10.79	21	65%	7.26

As described in the section on method, students in Class A struggled to attain the 80% reliability standard required to pass the proficiency test even after multiple attempts, necessitating a change in the way the test was administered. Some students in the class attempted the proficiency test many times more than the anticipated three attempts instructors believed it would take students to pass. Although not all students from Class A saved a copy of every attempt they made to pass the proficiency test, many students did save their final attempt before researchers terminated the test. For those students who attempted the test more than three times and who did save the results of their final attempt, the number of attempts they made and their corresponding reliability scores are recorded in Table 5. These scores show that some students attempted the proficiency test a high number of times before approaching the 80% reliability standard. This provides further evidence that, contrary to expectations, these students were not able to attain a high level of coding reliability.

Table 5

Class A Proficiency Test Scores Beyond Third Attempt

Final attempt recorded (if greater than 3)	Matching code labels	Overlapping time
4	71%	54%
4	80%	85%
4	91%	80%
4	95%	94%
4	96%	80%
5	65%	66%
5	100%	80%
6	95%	79%
8	89%	76%
9	60%	64%
12	55%	73%
12	84%	77%
14	75%	64%
14	81%	72%
15	89%	76%
15	99%	82%
28	91%	78%

The second finding from the proficiency test was that students showed improvement in their coding over the three attempts. As indicated in the Method section, I conducted a within groups ANOVA to provide an indication of whether or not the improvement in scores across three attempts on the proficiency test was statistically significant. The ANOVA showed a significant difference across the three proficiency test attempts (F= 24.492, p<.001). Pairwise comparisons revealed that there was a statistically significant difference between each of the three pairs of attempts (see Table 6). This finding indicates that there was statistically significant improvement in proficiency test scores with each successive attempt. This improvement was probably due to the effect of students receiving feedback on their previous attempt and then revisiting their coding to make corrections. This finding is not surprising, but it does give an indication that coding reliability can be learned, as will be discussed in the discussion section.

Table 6

Pairwise Comparison of Proficiency Test Attempts

Attempts Compared		Mean Difference	*SE*	*p*
1	2	-8.481	2.948	.024
2	3	-9.111	1.939	.001
1	3	-17.593	2.552	.001

The third finding from the proficiency test is that class membership was a significant factor influencing coding reliability scores. Students in Class A outperformed students in Class B on each of the three proficiency test attempts. An independent samples t-test showed that there was a statistically significant difference between the two classes on the first and third attempts (see Table 7). For the first attempt, there was a statistically significant difference for both the matching code labels scores and the overlapping time scores. For the third attempt, there was a statistically significant difference only for the matching code labels scores. The finding that the differences between scores from the two classes for at least some of the attempts were statistically significant indicates that class membership influenced coding reliability.

Table 7

Comparison of Proficiency Test Scores for the Two Classes

Attempt	Matching Code Labels		Overlapping Time	
	t	*p*	*t*	*p*
1	3.058*	.008	3.199*	.004
2	1.033	.315	.327	.747
3	2.842*	.010	1.986	.066

As indicated in the method section, I observed both classes prior to the administration of the proficiency test to monitor any potential differences between instruction for the two classes. I summarized my observations by listing the significant similarities and differences between the two classes that I believed might potentially influence the proficiency test scores.

Similarities between the two classes are as follows:

1. Both instructors explained the desired teaching competencies consistently, both using the handout listing the component parts of each competency as a guide.
2. Both instructors discussed the same readings and even used the same Powerpoint presentation.
3. Both instructors debriefed the first teaching experience as a class.
4. Both instructors gave the same hints to students about watching the expert teacher in the proficiency test video to determine whether the freezes are "all compliant" or "non-compliant." The instructors both said to watch the expert teacher to see if she was giving very specific verbal reinforcement, and if so then she was doing it to correct a non-compliant behavior. They gave similar hints about watching whether the expert teacher's movement in the gym was purposeful or not in order to discern whether it was teacher proximity or not.
5. Both instructors provided the same introduction to the computer and media labs.
6. Both instructors told students to start and stop codes about 2-3 seconds before and after the target video event.

Differences between the two classes are as follows:

1. Instructor B gave a more comprehensive introduction to the Studiocode software. She gave a systematic introduction to using the various functions of the software.
2. Instructor A had more time in the computer lab as a class and he was thus able to scaffold coding more. His students had almost an hour in class to work on the proficiency test, during which time they asked questions about how to code. Some of the questions that came up were when to start and stop a code and how to determine whether a freeze was "all compliant" or "non-compliant". Instructor B

also gave her students some time to start the proficiency test in class, but only for a few minutes.

3. Instructor A emphasized the importance of mastering "Teacher Talk" and had the students actually practice saying the words in class. In contrast, Instructor B tended to emphasize the importance of helping children develop an interest in lifelong fitness.
4. Instructor A tended to think about management in terms of control (i.e., Not "letting the kids get away with anything" and thinking about student misbehaviors in terms of a "power struggle"). Instructor B softened it a bit more by emphasizing the importance of learning children's names and not embarrassing them when implementing the discipline plan.
5. Instructor B referred the students to the desired competencies handout more than Instructor A.
6. Instructor A gave some of the rationale for the use of Studiocode: seeing is better than remembering.
7. When on the gym floor, Instructor A had the students practice teaching using the various desired competencies. In contrast, Instructor B almost always acted as the teacher modeling the competencies.
8. From the start of class, Instructor A introduced the graduate student researcher so that students were familiar with him and even asked him questions about the course. In contrast, Instructor B only introduced him when it was time to get signed consent from the students. This may have influenced students' willingness to engage with the study.

9. Instructor A used the actual proficiency test video in modeling some of the coding, but Instructor B used a different video of the same expert teacher. Instructor A also modeled coding three competencies (toe-to-toe, freeze, and long instructions) whereas Instructor B only modeled coding one competency (freeze). Instructor A also gave some extra comments to prime the students' noticing on the proficiency test. For example, he told the students to watch a particular child who would end up on time out later in the lesson.
10. Instructor A assigned his students to watch two extra videos where the expert teacher models the four-part PE lesson plan. Instructor B did not assign this, but modeled the four-part lesson plan on the gym floor in greater detail than Instructor A.
11. Instructor A's students did the proficiency test earlier in the semester than Instructor B's. By the time Instructor B did the proficiency test, her students had already taught in the schools a couple of times.

Despite the fact that this list of differences is more extensive than the list of similarities, it was my observation that both classes were remarkably similar in content, schedule, and teaching approach. I have merely reported on the differences in greater detail than the similarities because they help to explain the between-class differences in proficiency test scores.

During the follow-up interviews with the two instructors, I asked them to explain the difference in scores between the two classes. Both instructors commented that this difference could be explained, at least in part, by the amount of time that Instructor A had in the computer lab with the students during class. Because his class had almost twice as much time in the computer lab, the students had greater opportunities to ask him

questions and to work with one another on the proficiency test. The instructors also suggested that student motivation for Class A could have been higher because these students were told they had to attain 80% coding reliability before passing the proficiency test. In contrast, students in Class B might not have been as motivated to perform well on the proficiency test because they knew that even if they did not reach the 80% coding reliability, they only had to attempt the test three times. Despite this difference, Instructor B felt her students still would have been quite motivated to perform their best on the proficiency test. These comments from the instructors help to explain the observed difference in scores between the two classes.

Self-coding Audit Results

The results of the self-coding audit provide a contrasting perspective on the students' coding ability. For the self-coding audit, I audited 15 randomly selected codes from 20 students' coded videos of their own teaching. In total, I audited 300 codes, checking each instance to determine whether the student had used an appropriate code label to mark the video instance. Unlike the proficiency test that showed the students to be much less reliable coders than expected, the self-coding audit provides evidence that the students may indeed be reliable coders when subjected to a different measure.

Of the 300 randomly selected codes audited from 20 students, only 26 codes were used incorrectly. The mean reliability for these 20 students was 91% (SD = 0.06). This score exceeds the 80% reliability standard and aligns with the instructors' expectations for student coding ability. It does, however, demand explanation in light of the lower reliability scores found in the proficiency test.

Potential Sources of Variance in Reliability Scores

Interpreting the contradictory results from the proficiency test and the self-coding audit requires investigation into the potential sources of variance in these reliability scores. This section will attempt to explain the variance in the scores by first examining the specific codes that were problematic to students in the proficiency test and in the self-coding audit. Then it will present findings from student and instructor interviews that explain why these codes were problematic. Finally, these findings will be summarized into a theoretical model.

The number of errors for each code on the proficiency test was tracked for each attempt. When averaged for all participants and divided by the number of instances of each code in the expert coding, these values provide an indication of which codes were most problematic for students. Table 8 shows the error score for each code. Problematic codes have been indicated with an asterisk. Students struggled to code noncompliant freezes, all three labels for the instruction code, and three of the five labels for the discipline plan code. The least problematic codes had to do with the transition of equipment or students during a PE lesson.

During the self-coding audit, I noted the reasons for incorrect coding. Table 9 shows a brief description of these errors and how many times they occurred out of the 300 codes that I audited (there were only 26 incorrect codes out of the 300). The problematic codes for the self-coding audit were also problematic for the proficiency test, with the exception of two errors on a student transition code. The high frequency of errors on the first three codes in Table 9 (*Freeze – all compliant*, *Instruction - Short instructions*, and *Discipline plan – positive reinforcement*) is probably due to the fact that these are some of the most common teaching competencies and are used multiple times during a typical PE lesson.

Table 8

Average Error for Each Instance of Codes on Proficiency Test

Code	Attempt 1		Attempt 2		Attempt 3	
	Extra	Missing	Extra	Missing	Extra	Missing
ai: Freeze - All compliant	0.12	0.21	0.13	0.17	0.13	0.16
aii:Freeze - 1-4 noncompliant	0.71*	0.48*	0.64*	0.43*	0.58*	0.42*
bi: Transition - Equipment Onto Floor	0.03	0.17	0.06	0.09	0.03	0.06
bii: Transition - Equipment Off Floor	0.10	0.20	0.12	0.12	0.03	0.06
ci: Student Transition - Toe to Toe	0.00	0.02	0.08	0.00	0.02	0.00
di: Instruction - When before What	0.52*	0.09	0.51*	0.08	0.43*	0.05
dii: Instruction - Short Instruction	0.14	0.44*	0.20	0.17	0.14	0.12
diii: Instruction - long Instruction	0.42*	0.40*	0.47*	0.26	0.53*	0.17
ei: Discipline - Step 1: Pos. Reinforce.	0.11	0.57*	0.28	0.42*	0.27	0.28
eii: Discipline - Step 2: Proximity	0.47*	0.43*	0.48*	0.41*	0.50*	0.35
eiii: Discipline - Step 3: Quiet warning	0.05	0.20	0.06	0.16	0.08	0.14
eiv: Discipline - Step 4: 1st Timeout	0.00	0.07	0.02	0.05	0.08	0.08
ev: Discipline - Step 5: 2nd Timeout	0.04	0.48*	0.04	0.36	0.05	0.29

* Problematic codes indicated by average error score above .40

Table 9

Causes of Error in the Self-coding Audit

Coding error	Number of errors (out of 300 codes)
ai. Freeze - all compliant should have been *aii. 1-4 non-compliant*	10
dii. Instruction - Short instructions should have been *diii. Long instructions*	7
ei. Discipline plan - positive reinforcement that was actually only praise	4
ai. Freeze - all compliant that was not actually a freeze	1
cii. Student transition - splitting the class into two teams that was actually just lining up at the end of class	1
cii. Student transition - splitting the class into two teams that was actually the teacher going around the class dividing the students rather than using the specific technique for splitting the class specified by this code	1
eii.Discipline plan - teacher proximity that was really *eiii. Discipline plan - quiet warning*	1
eii. Discipline plan - teacher proximity that was not actually proximity	1

It is interesting to note that many of the errors in the self-coding audit involved using the wrong code when a different code was more appropriate (e.g., using *Short instructions* when the video segment called for *Long instructions*). This would indicate that errors in one code are connected to errors in another. This may imply that the proficiency test amplified the rate of error due to the way it was scored. For example, if a student were to code a video incident as *Short instructions* when the expert coded it as *Long instructions*, then the

proficiency test would make one deduction for having an extra *Short instructions* code and another deduction for missing a *Long instructions* code. Thus the proficiency test would count this error as two deductions whereas the self-coding audit would count it as only one. This implies that one of the potential sources of error in the coding reliability scores is the amplification of error caused by overly sensitive scoring.

The student and instructor interviews shed further light on the meaning of the low reliability scores from the coding proficiency test. These interviews revealed that despite the instructors' intent to make the coding scheme clearly defined and unambiguous, students still found multiple ways to interpret and implement the coding system. Thus the variance in reliability scores was due not only to differences in how experts and novices understood the desired teaching competencies but also to differences in how each coder implemented the coding system. The explanations given by students and instructors as to why the student reliability scores might have been so low on the proficiency test were classified based on whether they represent differences due to how students understood the teaching competencies or differences due to technical aspects of how students implemented the coding system. Table 10 shows the explanations that illustrate differences due to student understanding of the teaching competencies. Table 11 shows the explanations that illustrate differences due to implementation issues. Both tables include student and instructor comments that illustrate each explanation. The results of the interviews were most revealing. They suggest that the instructors' assumption that the coding system was unambiguous was erroneous. Students found many ambiguities in the actual implementation of the coding system. These were technical ambiguities that were typically quite trivial in terms of what they said about student understanding of the teaching competencies. For example, at the start of the expert teacher's PE lesson, she helped the students practice freezing in order to warm

Table 10

Follow-up Interview Explanations Dealing with Understanding of Teaching Competencies

Differences due to understanding of teaching competencies		
Explanation for low reliability	**Sample student comments**	**Instructor comments**
Distinguishing all compliant from 1-4 non-compliant for freezes	When the instructor modeled this one, he would say a clip was "non-compliant" and she couldn't tell why. Couldn't see students' misbehavior on the non-compliant ones.	Students probably are not picking up on them. He's not sure why this would be off as much. He points out that there are a lot more freezes in a lesson.
Positive reinforcement	Said she probably just wasn't paying attention to the positive reinforcement. Didn't know what should count as positive reinforcement when the teacher was going around the class talking to the students just a little bit. Didn't distinguish positive reinforcement from praise. She eventually "figured out" that the difference between whether it should count as positive reinforcement was whether or not the whole class was paying attention.	Pointed out the distinction between positive reinforcement and positive praise. Positive reinforcement is given to one child in order to help another student bring his/her behavior into compliance. Praise is simply given for the benefit of the compliant child. Students may not have picked up on this.
Teacher proximity	Said it was hard. She didn't do it herself when teaching and said she struggled with coding it in the expert video. The teacher didn't move very much, just a couple of steps if the student was close by, so it was hard to know whether to count it as proximity or not.	
Simultaneous or contiguous behaviors	She didn't know when to stop one code and start the other. For example, proximity and positive reinforcement flow into each other.	For the discipline plan, Instructor A talked about what he calls "behavioral chains". An example of a behavioral chain is the simultaneous act of scanning during a freeze, increasing teacher proximity to misbehaving students, providing positive reinforcement, giving instructions to the entire class, and then giving the warning to an individual student. For instructional purposes, these have been broken into discreet steps rather than the flow that naturally occurs.
Not perceiving the entire competency		The start and stop times for many of the codes was off. She says this is partly a nontrivial difference. Part of understanding each competency is learning to see the entire competency and not just capturing just the first couple of parts of the competency.

Table 11

Follow-up Interview Explanations Dealing with Implementation of Coding

Differences due to technical aspects of implementing the coding system

Explanation for low reliability	Sample student comments	Instructor comments
Unsure of appropriate length for coded clips	Thought her clips were too long, so she shortened them, which seemed to lower her scores drastically. Felt if she included all the criteria for each competency, then the clips would be way too long. Didn't think toe-to-toe should start with the freeze way before, even though the freeze was listed as one of the components of toe-to-toe. It seemed too long, plus it included a lot of other irrelevant codes that occurred between the freeze and the actual toe-to-toe pairing of students. Didn't pay any attention to the length of the clips.	Told students to make sure to include all the components of each competency, with a buffer of 1-2 seconds before and after.
How to code simultaneous, component, or successive codes	Simultaneous: Didn't realize on her first attempt that she could code more than one thing going on at a time. Component: Wasn't sure if it should be counted as when-before-what if it was also a part of a short instruction. She wondered which one to code or if she should code both. Successive: When the teacher gave several positive reinforcements in a row, she didn't know whether to code each individually or if she should lump them all together.	
Whether or not to code the startup of the lesson	At the start of the lesson, the teacher did some freezes to help the kids practice/warm-up. She coded this as "Freeze" but the expert coder did not count them.	
Short versus non instructions	Didn't realize all the really short things counted as short instructions, for example, "Put the beanbag on your head." Didn't know whether to include all the really short short instructions. She eventually thought that it was based on whether or not the kids were moving while the teacher spoke.	When the two instructors coded the video, they disagreed at first whether all the little teacher comments should count as short instructions. This occurred during the fitness routine when the teacher says something between each little activity. The instructors resolved this by deciding that it only counted as short instructions if the comments were directed to the entire class and not just a few individual children. Students probably saw these instructions as part of the overall fitness routine and not as instructions.

Table 11 (continued)

Follow-up Interview Explanations Dealing with Implementation of Coding

Differences due to technical aspects of implementing the coding system

Explanation for low reliability	Sample student comments	Instructor comments
Incomplete or flexible execution of a desired competency by the expert teacher	The expert teacher gave a warning but didn't say, "This is your warning." So the student was confused whether or not to code it as a warning since the instructors had specifically told them to say that. A boy got sent to time out three times, but the coding system only allowed for two time-outs so she didn't know how to code it.	
Student-teacher conference following a 2nd time out	The expert teacher sends a boy to time out and then holds an after class conference with him. One of the components of a second time out is this conference, but she was confused whether to code this as one or two separate codes and which labels to use. (the expert coders coded the conference as an independent "2nd time out")	
Short versus long instructions	Didn't realize at first that it had to be 60 seconds in order to count as long instructions	The difference between short and long instructions is whether it is under or over a minute. Students may not have picked up on that. It was a somewhat arbitrary number chosen to provide a clear cut-off to distinguish short from long instructions.
Video did not capture the entire behavior	Couldn't see the child to see if it was his first or second time out	Tried to give students hints to compensate for "off screen" behaviors. They told students to watch the expert teacher to see if she was doing something specific to correct a misbehavior, then it was a tell-tale sign there was a misbehavior.
Same code, multiple labels	When-before-what and short instructions are both labels on the instructions code. She wasn't sure how to label them since she wasn't supposed to double label them for this test.	For the sake of scoring the proficiency test, instructors told students not to attach more than one label to an individual code.

up the children and make sure they remembered the instructional expectations for the day. During coding, students were unsure whether to count these warm-ups as a freeze or not. Variance in how they coded this part of the lesson was not a valid measure of their understanding of the freeze teaching competency.

The instructors did not seem to be aware of the potential ambiguity of implementing the coding system. When asked to provide her overall interpretation of the proficiency test results, Instructor B attributed the low reliability scores to the difference between the ability of experts and novices to understand and perceive teaching competency. Similarly, Instructor A focused on explaining the difficulty of coding "behavioral chains." This is the term he used to talk about the simultaneous and interconnected nature of many of the teaching competencies (e.g., A teacher may be simultaneously giving instructions, increasing teacher proximity to a misbehaving child, and providing positive reinforcement to try to bring misbehavior into compliance). Instructor A attributed the low student reliability scores to the difficulty of coding behavioral chains. Essentially, he was making the same attribution as Instructor B: low student reliability was due to expert-novice differences in the understanding and perception of teaching competencies. I have chosen to call the type of variance in reliability scores addressed by the instructors the variance due to understanding of teaching competencies.

The instructors' attribution of low reliability scores to a lack of student understanding is not altogether invalid. Interview data from students also indicated that some of the variance in reliability scores was due to misunderstanding of the teaching competencies (see Table 10). For example, during the interviews one student said she simply did not understand teacher proximity very well. Other students indicated they simply were not paying very much attention to positive reinforcement while coding, which may indicate their

novice inability to attend to multiple and simultaneous aspects of instruction. What is interesting is that both instructors did not offer much in the way of explaining the ambiguity of implementing the coding system. This finding suggests they were not very aware of this other source of variance.

The instructors' inattention to variance due to implementation issues is somewhat surprising given the instructors themselves encountered some degree of implementation ambiguity when they each independently coded the expert video. For example, instructors themselves had differed in their use of the *Short Instructions* code. At one point in the expert video, the expert teacher has the children engaged in a fitness routine activity during which the expert stops the class frequently to show them the next fitness exercise. One of the instructors coded each of these instances as *Short Instructions* whereas the other instructor did not count them given their short length. Only through dialogue did the two instructors resolve this difference by agreeing to code as *Short Instructions* any teacher comments directed to the entire class. This example illustrates that even the two expert coders initially disagreed on at least some aspects of how to implement the coding system.

It is also interesting to note that neither students nor instructors are suggesting the technological ability of students to use the Studiocode software might be an explanation for variance. The technological knowledge threshold for learning to use Studiocode seems to be very low. The only mention of technological ability interfering with coding scores was made by Instructor B when referring to a student in her course who she described as being technologically challenged in general. She said that this student struggled not only with Studiocode but with computers in general and that might have contributed to his lower scores. Instructor B implied that the student was an outlier in terms of technological ability. The fact that instructors and students did not mention technological ability as a factor

influencing coding scores can be interpreted to mean that it was not a significant source of variance in the reliability scores.

Given the foregoing investigation into the potential sources of variance in the reliability scores, a model can be used to summarize the findings. The total variance in the proficiency test reliability score is represented by the formula in Figure 7.

$$V_T = V_U + V_I + V_S$$

V_T = Total Variance in the Reliability Scores
V_U = Variance Due to Understanding of Teaching Competencies
V_I = Variance Due to Implementation of Coding System
Vs = Variance Due to Scoring (e.g. Amplification of error)

Figure 7. Total variance in coding reliability scores.

At the outset of this study, the assumption made by researchers and instructors was that the reliability of student's coding scores would be a basis for making inferences about the students' understanding of teaching competencies. However, given that the total variance in reliability scores is composite, the use of reliability scores to make inferences about students' understanding of teaching competencies is only valid if the other sources of variance are minimized. This formula indicates that even though students may not have obtained very high reliability scores on the proficiency test, they still may have had a fairly good understanding of the teaching competencies if a large proportion of the variance was due to V_S or V_I.

Comparing Variance in the Proficiency Test and the Self-coding Audit

This analysis of the sources of variance in the reliability scores allows us to propose a resolution to the contradictory results of the proficiency test and the self-coding audit. One

possible explanation is that the proficiency test scores included a greater amount of variance due to scoring and variance due to the implementation of the coding system than the self-coding audit. This explanation is supported by the fact that the scoring of the proficiency test was automated in a way that did not make allowances for individual interpretation of the implementation of the coding system.

For example, it is not uncommon for a PE teacher to use several statements of positive reinforcement in a row, directed at different children. During coding, students would be faced with the dilemma of deciding whether to code each of these statements as an independent instance of *Positive Reinforcement* or whether to lump them all together into one instance. Not knowing that the expert coder chose to lump these statements all together, the students who code each individual statement will receive significant deductions to their coding score. With the automated system of scoring for the proficiency test, coding each statement individually might result in multiple deductions, one for each "extra" code. In contrast, because the self-coding audit was conducted by a researcher examining each individual code in context, the researcher could make the judgment call that this student was using the *Positive Reinforcement* code accurately. Thus the self-coding audit was able to eliminate variance due to differences in how the students implemented the coding system.

As explained earlier, the self-coding audit was also less prone to introduce variance due to the amplification of error in scoring. Thus the self-coding audit was likely to provide a more valid base for making inferences about student's actual understanding of teaching competencies (see Figure 8).

Despite this argument for preferring the self-coding audit score as a more trustworthy reflection of actual student understanding of the teaching competencies, there are several other factors that make the proficiency test and the self-coding audit

incommensurate. First, the proficiency test detected the identification of false positive instances of each teaching behavior. For example, if the teacher gives positive praise in the video, then the proficiency test coder will either incorrectly identify it as positive reinforcement or correctly not code it as positive reinforcement. In the self-coding audit, the fact that the coder did not code this event as positive reinforcement does not necessarily mean the coder understands the difference between positive reinforcement and positive praise. Because the coding was not exhaustive for the self-coding audit, the coder may simply have ignored this instance. Thus the proficiency test was arguably a more sensitive measure of actual understanding of the teaching competencies [i.e. V_U (Self-coding audit) $\neq$ V_U (Proficiency test].

If: V_I(Self-coding audit) $<$ V_I(Proficiency test)

And: Vs(Self-coding audit) $<$ Vs(Proficiency test)

And: V_U (Self-coding audit) $\sim$ V_U (Proficiency test)

And: $V_T = V_U + V_I + V_S$

Then: V_T (Self-coding audit) is a better approximation of V_U (Actual) than V_T (Self-coding audit).

Figure 8. Rationale for preferring the self-coding audit score to the proficiency test score.

Second, there may be other effects on coding due to the differences between content of the videos. Whereas for the proficiency test, students coded videos of an expert teacher, for the self-coding audit, students coded videos of themselves teaching. Coding oneself may be more difficult, because it is less likely that novice teachers will fully execute each teaching competency and thus there will be more ambiguous instances

of teaching behavior to try to fit within the coding scheme. On the other hand, coding oneself may be easier because student teachers will recall their intentions to implement the desired teaching competencies during the lessons captured on video. Thus when students code their own video they will already know whether they should "count" each instance of teaching behavior.

Third, the scores on the proficiency test and self-coding audit may have been subject to the effect of practice. The proficiency test was administered before students had extensive experience using Studiocode. The self-coding audit, however, was conducted at the end of the course. By that point, students had already practiced using Studiocode on the proficiency test and for three of their own teaching videos.

Given these arguments, it is difficult to conclude which of the two scores is a more valid reflection of actual student understanding of the teaching competencies. Given this impasse, it is necessary to report that the findings from this study are inconclusive about the reliability of student coding in this course.

Chapter 5: Discussion

The findings from this study begin to define the issues associated with video coding reliability. In this section I will explore some of the implications of the findings. First, I will consider whether or not coding reliability can be used as a basis for making inferences about students' understanding of teaching competencies. Second, I will discuss how findings from this study show that coding reliably is something that can be learned and taught. Third, I will consider the tradeoffs involved in using a coding system for assessment in teacher education. In conclusion, I will outline the contributions of this study to the literature on video analysis tools and potential questions for future research.

Using Coding Reliability Scores to Make Inferences about Understanding

Although not all of the variance in the reliability scores was due to actual differences in how the students understood the teaching competencies, the findings from this study indicate that coding reliability can still be used profitably to identify student misunderstandings. When the problematic codes from the proficiency test are compared to the problematic codes in the self-coding audit and are contextualized by follow-up interview data, there is a convergence of findings. Some of the same codes that had low reliability scores on the proficiency test were also identified as problematic by these other measures. Specifically, students struggled with the four following concepts:

1. Teacher proximity
2. Distinguishing positive reinforcement from positive praise
3. Noticing non-compliant behavior during a freeze
4. Distinguishing short and long instructions.

These four aspects of the coding system are non-trivial, meaning understanding them is important to developing teaching expertise. Thus the reliability scores provide some

information about how well students understand the teaching competencies represented by the coding scheme.

That being said, the value of coding reliability scores for diagnostic purposes was severely limited in this study by the high amount of variance in the scores that was due to extraneous factors not related to the students' level of understanding. As explained in the Findings section, one way to represent this situation is by the formula: $V_T = V_U + V_I + V_S$, where V_I and V_S represent variance due to extraneous factors (V_I = Variance due to implementation of coding system, V_S= Variance due to the scoring of coding). The results of this study provide evidence that V_I and V_S were high. Some of the codes identified as problematic by the proficiency test were not identified as problematic in the self-coding audit (e.g., *ev. Discipline plan – Step 5: Second timeout* and *di. Instruction – When before what*). If student understanding of the codes was fairly stable between these two measures, then it would be expected that the same codes would be problematic in both measures of coding ability. However, since some codes were not identified as problematic by both measures, there is good reason to believe at least some of this variance was due to extraneous factors. Furthermore, the follow-up interviews provided numerous explanations of how such variance could be caused by differences in technical aspects of how the coding system was implemented (V_I) or scored (V_S). The fact that the reliability scores obtained in the proficiency test vary so greatly from the scores obtained in the self-coding audit also casts suspicion on the ability to use reliability scores to make inferences about students' understanding.

The only finding from this study that points to the validity of these reliability scores is the apparent high degree of expert inter-coder reliability. At the outset of the study, the two instructors independently coded the expert video. Following this

independent coding, the instructors engaged in dialogue to resolve discrepancies between their coding. The result of this dialogue was a joint coding that represented the expert coding standard. A comparison of Instructor B's independent coding against the joint coding standard showed a high degree of reliability. This finding could be interpreted to mean that variance in reliability was due mainly to whether or not the coder had an expert understanding of the teaching competencies and that other sources of variance are negligible.

This interpretation, however, conflicts with the body of evidence already cited that supports the explanation that the reliability scores in this study included a great deal of variance due to extraneous sources. The high degree of expert inter-coder reliability can be reconciled with the other findings from this study by taking into account several factors.

First, the initial independent codings by the two instructors may not have attained such a high degree of reliability. Because Instructor A's independent coding was lost, the actual initial expert inter-coder reliability score is unknown. We do know, however, that the instructors disagreed on the implementation of at least two codes: *ev. Discipline Plan – Step 5: Second timeout* and *dii. Instruction – Short instructions*. Their disagreement on these codes was not about the meaning of the teaching competencies but had to do with the more trivial aspects of how to implement the coding system.

Second, the instructors may not have had as much variance in how they implemented the coding system due to their experience working together. These teachers had collaborated in their teaching of the course and use of Studiocode for several semesters prior to this study. It is not unlikely that through this interaction, they had already arrived at a shared understanding of how to implement most of the coding

system. Thus we might not expect them to differ much in how they implemented the coding system.

Third, the amount of variance between two scores is likely to be less than the variance between many scores. The expert inter-coder reliability score was based only on the scores of the two instructors. The reliability scores obtained from the students was based on the experiences of 49 students implementing a coding system. It is likely that those 49 students varied to a greater degree than the two instructors in how they implemented the coding system. If 49 instructors were to code the same video, there would likely be more variance in how they implemented the coding system.

Given these considerations, it is reasonable to accept the interpretation that the reliability scores in this study were subject to a high degree of variance due to extraneous factors (i.e., V_I and V_S). This finding has implications for those researchers interested in measuring coding ability. The maxim that reliability does not ensure validity needs to be considered in the measurement of conceptual understanding as reflected by coding ability.

Learnability/Teachability of Coding Reliably

Findings from this study also suggest that coding reliably is something that can be learned and taught. This study showed that students improved their coding reliability across the three attempts on the proficiency test. The only formal feedback students received between attempts was an indication of how many instances of each code were missing and extra. During follow-up interviews, some students indicated that this feedback was not very useful to them. They described the situation where they knew they were missing several instances of a particular code, but after reviewing the video and their coding they still did not know where those missing instances were located. One

student explained that this was so frustrating that she began a systematic approach to making one change at a time and then re-scoring her proficiency test. These comments indicate that students would have benefited from more specific feedback. What is interesting is that despite the lack of very specific feedback, students still showed significant improvement across the three attempts (see Table 12). This suggests that coding reliably is learnable.

Table 12

Improvement in Scores Across the Three Proficiency Test Attempts

	Attempt 1 (Mean score)	Attempt 2 (Mean score)	Attempt 3 (Mean score)
Matching Code	49%	57%	63%
Overlapping Time	55%	65%	68%

That coding reliably is teachable is suggested by the difference in scores between Class A and Class B. On the proficiency test, Class A consistently outperformed Class B. This difference may be explained by the amount of social scaffolding students had access to in the two classes during the proficiency test. One of the main differences between the two classes was the amount of in-class time students had to work on the proficiency test. Students in Class A had more in-class time in the computer lab where they worked on the proficiency test. The researcher who was observing both classes noted that during this in-class time students tended to ask each other and the instructor questions about the implementation of the coding (e.g., how to decide when to start and stop a code) and about the meaning of the codes (e.g., how to decide if something is really positive reinforcement or just praise). It may be that this social scaffolding provided opportunities for students to learn coding skills from one another.

Further evidence that coding reliably is something that can be learned and taught is that students improved their coding reliability in a very short time frame. The context of this study must be kept in mind. These students were in a block class, meaning the course lasted only half a traditional semester or about 8 weeks. That meant that students only had two weeks of class prior to their entering the local elementary schools to begin their field teaching experiences. During those two weeks, meeting only twice a week, students had to learn about each of the fifteen teaching competencies. Furthermore, during those first two weeks students had to learn and prepare for other aspects of teaching—things like the four components of an effective PE lesson plan, the guiding principles for PE teaching, and the preparation of teaching materials such as instructional signs and timed music CD's. On top of these demands, students learned to navigate the Studiocode software. In short, prior to the proficiency test, these students had very little instructional time dedicated to learning coding (about 5-8 hours). Despite these limitations on students' time, students were still able to demonstrate moderate coding ability on the proficiency test and improvement over the three testing attempts.

These considerations may indicate that given more time to practice coding, students may become quite proficient coders in a relatively short amount of time. For this class, coding had to be learned quickly to facilitate learning for one academic block. At BYU there are students majoring in PE teaching. These students have the opportunity to use Studiocode over the course of several semesters. Given that amount of exposure, these students would be likely to become quite reliable coders. The implication for instructors desiring to use a coding system in teacher education is clear: learning to code will take some time. Although for a relatively straightforward and highly operationalized coding system such as that used in this study the amount of time needed to train students

in coding may be very small, students still need the opportunity to practice and receive feedback on their coding.

Those instructors or researchers interested in helping students to become proficient coders will also need to consider how they can support the social learning of coding skills. This study has suggested that a great deal of variance can be introduced into reliability scores by the technical implementation of a coding system. In order to minimize the amount of the variance due to implementation, instructors and researchers will need to ensure that all students/participants engage in dialogue to arrive at a shared understanding of implementation issues.

Tradeoffs in Instructional Use of Coding Systems

Despite the fact that students failed to reach the expected level of coding competency during the proficiency test, both instructors remained optimistic about the instructional value of the coding system used in the course. During their follow-up interviews, they expressed the belief that even though students did not demonstrate expert coding ability, the coding system still provided the students with the information they needed in order to improve their teaching practice. The instructors argued that through the iterative process of teaching a PE lesson, coding that lesson, and then teaching another PE lesson, students learned to filter through their teaching performances to identify areas for improvement. Because students knew that at the end of the course they would need to submit their best example of each of the fifteen teaching competencies, students would, after each coding session, naturally focus on those teaching competencies they had not yet mastered in previous PE lessons. Students may not have been able to identify all teaching competencies with an expert eye, the instructors argued, but for beginning teachers that level of coding reliability really may not be necessary.

The instructors' response to the findings of this study point to the fact that there are many tradeoffs in the use of coding systems and these must be balanced in light of the instructional objectives. Like more traditional forms of assessment, coding can be evaluated based on the tradeoffs between various dimensions of the assessment: formative versus summative, objective versus subjective, ease of scoring, time-intensiveness for students, ease of construction, content coverage, generative versus recognition, etc.

There are often trade-offs among these dimensions of assessment. For example, essay questions in a traditional paper-pencil exam can be very valid reflections of what students actually understand; however, they come at a rather high cost in terms of difficulty of scoring, subjectivity, time-intensiveness and content coverage. On the other hand, multiple-choice test items, when properly constructed, can be quite objective, easily graded, and cover a lot of content in a short amount of time, but it is very difficult to construct effective multiple-choice items that measure higher-level learning outcomes.

The coding system considered in this study could be characterized in the following way. It tended to be recognition-based rather than generative because students used the codes to identify instances of teaching competencies. If the students had annotated their coded video with textual comments, then the coding system might have been more generative. Studiocode provides the functionality needed for textual annotation, but for this iteration of the course instructors did not ask students to use this function extensively. Code annotations can be very time intensive for students to complete and time intensive for instructors to grade. Furthermore, code annotations introduce a greater degree of subjectivity into the coding assessment. At the same time code annotations could allow students to explore the nuances of teaching in a dynamic

classroom environment. Given that this course was for beginning teachers, the instructors chose to focus student attention on the mastery of a few well-defined teaching competencies. Instructors felt these students needed to develop a foundation of teaching skills to help them be successful before the students began exploring the nuances of teaching and developing their own personalized teaching style. Thus they opted for a recognition-based, objective coding system.

The coding system used in EXSC 375 is intended to be a formative assessment because it is based on the assumption that when students are asked to look for the best examples of target teaching competencies in the videos of their own teaching, they will notice the competencies for which they do not yet have good examples. This focuses students' teaching practice on the competencies they have not yet mastered. This coding system could play a stronger formative role if it helped students to explicitly describe the quality of each instance of a teaching competency. One way to do this would be to ask students to annotate codes with evaluative comments. The degree to which students could accurately evaluate their coded instances would be an interesting area of future research about the use of video analysis tools in teacher education.

In a way, the coding system used in this course is more explicitly summative than formative. Students were graded based on their demonstration that they could competently execute each of the desired competencies. In other words, they were graded based on whether or not they could use the coding system to identify instances of each competency in their own teaching videos. Students did not receive external feedback on their coding from an instructor or peers which might have helped them to improve their coding abilities and their understanding of the teaching competencies. One of the reasons instructors did not provide this feedback may have been that doing so would have been

time-intensive and thus logistically impossible. Indeed, one of the benefits of using the coding system is that it allows for students to self-edit their own videos so that instructors can focus on those video clips that students themselves have selected for evaluation. Another reason instructors may have chosen not to evaluate interim coding might have been to allow students the psychological space to make mistakes as they practiced executing each of the teaching competencies.

Contribution and Future Research Questions

As discussed in the literature review, research has established the benefits of using video in teacher education. Research has also demonstrated the importance of scaffolding video viewing in order for students to gain more from video. An emerging set of video analysis software tools shows promise for the scaffolding of student video viewing by helping students code videos along dimensions of expert teaching. Very little research has documented the use of these tools, particularly as they pertain to the evaluation of pre-service field teaching experiences. This study begins to systematically investigate the potential of video coding by considering the reliability of student coding as compared to expert coding.

The original intent of this study was to demonstrate whether students could reliably code videos of their own teaching when given a well-defined coding scheme. If researchers could demonstrate that students could be trained to be reliable coders, then researchers could begin to consider whether or not this coding provided the kind of feedback needed for students to engage in more meaningful teaching practice and ultimately to become better teachers. The simple answer to the research question is that students demonstrated moderate coding reliability when coding a video of an expert teacher, much lower than was anticipated by myself and the instructors who had expected students to attain 80% reliability.

This finding, however, is called into question when triangulated with other data from this study. Although the proficiency test showed that students were only moderately reliable coders, data from the self-coding audit and follow-up interviews call into question the meaning of these reliability scores. The self-coding audit showed that when coding reliability was measured in a way that only took into account students' ability to identify positive instances of codes, they attained a much higher level of reliability. The follow-up interviews revealed that a large proportion of the variance in the proficiency test reliability scores may be explainable in terms of extraneous factors, such as the technical ambiguities of implementing a coding system or the way coding was scored. If the reliability scores include a large amount of extraneous variance that does not reflect actual student understanding, then it is possible that each student's understanding of the teaching competencies represented by the coding scheme is greater than the proficiency test results would indicate. This would indicate that coding reliability may not be a sound basis for making inferences about understanding of teaching competencies.

To conclusively answer the question of whether or not these students could be trained to be reliable coders, researchers would need to repeat this study with modifications to minimize the amount of extraneous variance in the reliability scores. The findings from this study indicated that there were two sources of extraneous variance. First, there was variance due to technical differences in how students implemented the coding system (V_I). In a follow-up study, researchers would need to work with instructors to ensure that when students are trained in coding that the technical ambiguities of implementing the coding system are addressed. Findings from the follow-up interviews (see Table 11) identify several of those technical implementation ambiguities and provide a good indicator of the types of ambiguities that are likely to lead to implementation variance. Given that this study shows

students' coding ability can improve with practice and minimal feedback, the prospect of successfully reducing V_I through improved training is promising.

Second, a follow-up study would need to address the extraneous variance due to the way in which coding reliability is scored (V_S). In this study we found that deducting for both missing and extra instances of codes may have magnified the amount of error in reliability scores (i.e., some errors were double counted). This problem may be due to the unit of analysis used in scoring the coding. Whereas the proficiency test used each code as a unit of analysis, future researchers may want to use video events as the unit of analysis. The high level of reliability found in the self-coding audit indicates that changing the way coding is scored may lead to reliability scores that are truer reflections of actual student understanding.

The instructors' reaction to the low reliability scores found in this study is interesting. Given the high level of student frustration with the proficiency test, the instructors' perceptions that it contributed relatively little to student learning, and the limited amount of instructional time in this course, the instructors plan to discontinue this part of instruction. Rather than eliminating the requirement for students to analyze video of an expert teacher, the instructors might want to try providing students with their own coding with annotations. This would provide a scaffold for student learning without being too time intensive.

That the instructors plan to discontinue the proficiency test is not very surprising. The provocative part of their response to this study is their intention to continue the self-coding activity. The instructors indicated that despite the low reliability found in this study, they still perceived the self-coding activity to be productive for the students and intend to continue using it in the future for the course. This reaction raises a key question—if student coding is only moderately reliable, at least when compared to an expert standard of coding, can it still be valid enough to provide self-feedback that improves practice?—or, in other

words—how reliable is reliable enough? These instructors indicated that even though students could not reliably code like an expert, they could still obtain valuable insights into their own teaching practice that resulted in improved teaching practice.

Future research might profitably investigate the types of insights students gain into their own teaching practice through their coding activities and how those insights influence their subsequent field teaching experiences. It might be interesting to compare what novice teachers believe they need to change about their teaching after coding video of their own teaching with what experts believe they need to change about their teaching after coding the same video. Future research might also investigate how textual annotation of video coding influences the quality of feedback that students generate through their coding activities.

References

Berliner, D. (1992). The nature of expertise in teaching. In A. D. F. Oser, & J. Patry (Ed.), *Effective and responsible teaching* (pp. 227-248). San Francisco: Jossey-Bass.

Bitner, T., & Kratzner, R. (1995). *A primer on building teacher evaluation instruments*. Paper presented at the Midwest Educational Research Association.

Bransford, J. D., Barron, B., Pea, R. D., Meltzoff, A., Kuhl, P., Bell, P., Stevens, R., Schwartz, D. L., Vye, N., Reeves, B., Roschelle, J., & Sabelli, N. H. (2006). Foundations and opportunities for an interdisciplinary science of learning. In K. R. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences*. Cambridge: Cambridge University Press.

Brawdy, P., & Byra, M. (1994). *A comparison of two supervisory models in a preservice teaching practicum*. Paper presented at the American Educational Research Association.

Brophy, J. (2004). Introduction. In *Using video in teacher education* (Vol. 10, pp. ix-xxiv). Oxford, UK: Elsevier Ltd.

Bullough, R. V., Young, J., Erickson, L., Birrell, J. R., Clark, D. C., & Egan, M. W. (2002). Rethinking field experience: Partnership teaching versus single-placement teaching. *Journal of Teacher Education, 53*(1), 68-80.

Chan, P. Y. K., & Harriss, R. C. (2005). Video ethnography and teachers' cognitive activities. In J. Brophy, & Pinnegar, S. (Eds.) (Ed.), *Learning from research on teaching: Perspective, methodology, and representation* (Vol. 11, pp. 337-375). Oxford, UK: Elsevier Ltd.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 33-81.

Chen, W. (2002). Six expert and student teachers' views and implementation of constructivist teaching using a movement approach to physical education. *The Elementary School Journal, 102*, 255-272.

Cherry, G., Fournier, J., & Stevens, R. (2003). Using a digital video annotation tool to teach dance composition. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 5*(1).

Collins, J. L., Cook-Cottone, C. P., Robinson, J. S. & Sullivan, R. R. (2004-2005). Technology and new directions in professional development: Applications of digitial video, peer review, and self-reflection. *Journal of Educational Technology Systems, 33*(2), 131-146.

Cunningham, A. B., S. (2002). *Using digital video tools to promote reflective practice.* Paper presented at the Society for Information Technology and Teacher Education International Conference, Chesapeake, VA.

Everhart, B., Johnson, R., & Brantley, B. (1996). *Preservice teaching in controlled and natural settings prior to student teaching in physical education*. Paper presented at the Annual Meeting of the American Educational Research Association.

Goldman-Segall, R. (1994). Challenges facing researchers using multimedia data: Tools for layering significance. *Computer Graphics, 28*(1), 48-51.

Goodwin, C. (1994). Professional vision. *American Anthropologists, 96*(3), 606-633.

Hiebert, J., Gallimore, R., & Stigler, J. W. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher, 31*(5), 3-15.

Hogan, T., Rabinowitz, M., & Craven, J. A. (2003). Representation in teaching: Inferences from research of expert and novice teachers. *Educational Psychologist, 38*, 235-247.

Jensen, R. A., Shepston, T. J., Connor, K., & Killmer, N. (1994). *Fear of the known: Using audio-visual technology as a tool for reflection in teacher education*. Paper presented at the Annual Meeting of the Association of Teacher Educators.

Jones, L. M., O. (2004). The possibilities and constraints of multimedia as a basis for critical reflection. *Cambridge Journal of Education, 34*(3), 279-296.

Karppinen, P. (2005). Meaningful learning with digital and online videos: Theoretical perspectives. *AACE Journal, 13*(3), 233-250.

Kipp, M. (2001). ANVIL: A generic annotation tool for multimodal dialogue. *Eurospeech*.

Max Planck Institute for Psycholinguisitics: General description Media Tagger.

Nemirovsky, R., Dimattia, C., Ribeiro, B., & Lara-Meloy, T. (2005). Talking about teaching episodes. *Journal of Mathematics Teacher Education, 8*, 363-392.

Nicol, C., & Crespo, S. (2004). *Learning to see in mathematics classrooms.* Paper presented at the 28th Conference of the International Group for the Psychology of Mathematics Education.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

O'Donnell, A. M., Reeve, J., & Smith, J. K. (2007). *Educational Psychology: Reflection for Action*. Hoboken, NJ: John Wiley & Sons, Inc.

Pailliotet, A. W. (1995). I never saw that before: A deeper view of video analysis in teacher education. *Teacher Educator, 31*(2), 138-156.

Pea, R., & Hay, K. (2002). *Report to the National Science Foundation: CILT Workshop on Digital Video Inquiry in Learning and Education*. Stanford University, Stanford, CA.

Pecheone, R. L., Pigg, M. J., Chung, R. R., & Souviney, R. J. (2005). Performance assessment and electronic portfolios: Their effect on teacher learning and education. *The Clearing House, 78*(4), 164-176.

Poole, J., & Wessner, J. (2003). *The transition from student to teacher: Developing a self-assessment culture for professionalism in teacher preparation programs*. Paper presented at the PAC-TE Teacher Education Assembly.

Preston, M. D., Campbell, G. A., Ginsburg, H. P., Sommer, P. & Moretti, F. A. (2005, July 2, 2005). *Developing new tools for video analysis and communication to promote critical thinking.* Paper presented at the Ed-Media: World Conference on Educational Multimedia, Hypermedia, and Telecommunications, Montreal, Canada.

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal, 28*(1), 63-88.

Schmidt, C. P., & McCutcheon, J. W. (1994). Verbal versus nonverbal cues in evaluations of teaching. *Journal of research and development in education 27*(2), 188-125.

Schon, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Sharpe, T. (1997). Using technology in preservice teacher supervision. *Physical Educator, 54*(1), 11-19.

Sherin, M. G. (2004). New perspectives on the role of video in teacher education. In J. Brophy (Ed.), *Using video in teacher education* (Vol. 10, pp. 1-27). Oxford, UK: Elsevier Ltd.

Sherin, M. G., & van Es, E. A. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education, 13*(3), 475-491.

Shulman, L. (1992). Toward a Pedagogy of Cases. In J. H. Shulman (Ed.), *Case Methods for Teacher Education* (pp. 1-30). New York: Teacher College Press.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Stevens, R. (2001). Video Traces: Media Rich Annotations for Learning and Teaching. University of Washington.

Stevens, R., & Toro-Martell, S. (2003). Leaving a trace:Supporting museum visitor interaction and interpretation with digital media annotation systems. *The Journal of Museum Education, 28*(2).

Struyk, L. R., & McCoy, L. H. (1993). Pre-service teachers' use of videotape for self-evaluation. *Clearing House, 67*(1), 31-34.

Thomson, W. S. (1992). Using videotape as a supplement to traditional student teacher supervision.

Vanderbilt, C. a. T. G. a. (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher, 19*(5), 2-10.

Wang, J., & Hartley, K. (2003). Video technology as a support for teacher education reform. *Journal of Technology and Teacher Education, 11*(1), 105-138.

West, R. E., Rich, P., Shepherd, C., Recesso, A., & Hannafin, M. (2007). *Supporting induction teachers' development by examining video evidence of teaching performance*. Paper presented at the American Education Research Association.

Winograd, K., Higgins, K., McEwan, B., & Haddon, L. (1995). *Relinquishing field supervision to the mentors: University teacher educators rethink their practice.* Paper presented at the Annual Conference of the American Educational Research Association.

Zimmerman, B. J. (2001). *Self-regulated Learning adn Academic Achievement: Theoretical Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.