Teacher Ratings of Student Engagement with Educational Software: An Exploratory Study

□ Robert L. Bangert-Drowns
Curtis Pyke

The quality of students' learning engagement may significantly influence their learning. Can teachers accurately judge student learning engagement with educational software? In this exploratory study, 3 fifth-grade teachers used a seven-level taxonomy to rate the frequency of different forms of engagement among 42 students interacting with different types of educational software. Teachers spontaneously treated the seven levels of engagement as a continuum, rating students highest on one level or a set of contiguous levels. Teachers generally agreed when ranking students by their typical levels of engagement, but disagreed regarding the actual frequencies of different engagement types. Ratings of software engagement conceived of as interpretive activity were correlated significantly with student reading test scores. Given the authentic classroom conditions in which this study took place, the results are promising for the classroom utility of the seven-level conception of student engagement with software.

□ Students are often enthusiastic and persistent in their interactions with educational software. But engaged students interact with software in qualitatively distinct ways. Some work independently, strategically, and creatively. Others depend on clear directions. Still others move from task to task without apparent plan. One might expect very different learning effects from these different styles of engagement. Teachers who make educational use of computer software need to distinguish different qualities of student engagement, so they can better anticipate and respond to different qualities of student learning. Teachers also could aid researchers evaluating the efficacy of educational software in light of learner engagement. This study explores teachers' abilities and difficulties in evaluating software engagement in conventional classrooms.

Learning Engagement

For most educators and researchers, engaged learners are thought to be more intensively and extensively involved—behaviorally, intellectually, and emotionally—in their learning tasks, than those who are not. Engagement entails instrinsically motivated involvement of integrated cognitive processes: creating, problem-solving, reasoning, decision-making, and evaluation (Kearsley & Shneiderman, 1998). Bangert-Drowns and Pyke (2001) called engagement "the mobilization of cognitive, affective, and motivational strategies for interpretive transactions" (p. 215).

However, researchers rarely operationalize engagement as a multidimensional phenomenon. Some researchers conceptualize engagement as "time-on-task" (e.g., Kumar, 1991; Martens, Bradley, & Eckert, 1997) or intrinsic motivation (e.g., Skinner & Belmont, 1993). Some measure engagement as self-regulated learning (e.g., Ryan & Patrick, 2001). Some researchers define engagement in terms of the characteristics of the students' learning environments, such as the quality of classroom dialogue (Nystrand & Gamoran, 1991) or the culture of the college campus (Kuh, 2000). Each of these captures crucial aspects of engagement, but none helps distinguish qualitatively different forms of engagement in classroom learning.

Corno and Mandinach (1983) made an early articulation of distinguishable forms of learning engagement. They imagined well-developed self-regulated learning as the fullest manifestation of engagement, coordinating sets of knowledge acquisition and knowledge transformation skills. Recipient learning, by contrast, involved impoverished acquisitional and transformational skills. Resource management is engagement with relatively strong knowledge-acquisition strategies; task-focused learning is engagement with relatively strong transformational strategies. Learners may employ different forms of engagement depending on their abilities and the learning situation.

Other researchers articulated multidimensional views of engagement. Nystrand and Gamoran (1991) distinguished disengagement, procedural engagement (minimal attention to task demands), and substantive engagement (sustained academic commitment). Lee and Anderson (1993), in a scheme for coding transcripts of students' classroom behavior, distinguished seven levels of engagement by kinds of behavioral, cognitive, and self-initiated involvements in learning. Ainley (1993) identified six types of engagement in a cluster analysis of measures of students' general ability and learning goals and beliefs. Four groups of students showed problematic involvement in learning, one group manifested learning that was compliant with specified learning goals, and one group showed high commitment to learning. Though these schemes were developed independently, they identify progressively more sophisticated levels of engagement, levels that entail more sustained effort and more strategic deployment of cognitive processes. However, none of these theoretical frameworks or coding rubrics was designed for teacher use in classroom instruction. And none is directly useful in elucidating student engagement with educational software.

Engagement in Literacy and Problem Solving

To the degree that sophisticated engagement entails self-regulated learning, it entails problem solving. Successfully engaged learners identify learning goals, deploy strategies to bridge the problem space, monitor progress toward goals, and adapt their strategies. Similarly, learning engagement in schools is often a literate act, an encounter with an organized body of knowledge that must be decoded, interpreted, and integrated in meaning-making processes.

This intimacy among engagement, problem-solving, and literacy is often reflected in the research on engagement. Guthrie (1996), for example, described engaged acts of literacy as "possessing a variety of motivations to gain conceptual understanding by using cognitive competencies and participating in a diversity of social interchanges" (p. 434), a definition consistent with learning engagement in general. Similarly, Yang (2002) studied engagement of students as they synthesized discourse in a hypermedia learning environment. She identified 32 interrelated metacognitive, cognitive, and affective processes that characterized engagement in a self-regulatory task that combined learning, literacy, and problem-solving.

Given engagement's relatedness with "literate," interpretive acts, conceptualizations of "high literacy" might extend one's understanding of engagement. Certainly, some researchers' definitions of literacy are closely entwined with definitions of engagement (e.g., Guthrie, 1996). Bereiter and Scardamalia (1987; Scardamalia, Bereiter, McLean, Swallow, & Woodruff, 1989) define *high literacy* as a strategic, intentional, reflective effort to satisfy curiosity in knowledge-transforming ways (i.e., intrinsically motivated self-regulation).

But other conceptions of high literacy add new possibilities to the notion of engagement.

Cognitive flexibility entails spontaneous, adaptive restructuring of schemas to solve problems of comprehension and application (Jacobson & Spiro, 1995; Spiro & Jehng, 1990). Cognitive flexibility brings active knowledge to the interpretation of novel cases and can be fostered by interactions with specially designed hypertexts that allow users to review electronic texts in varied combinations of predefined overarching themes. *Aesthetic reading,* instead of emphasizing the content of the text itself, makes the emotional and intellectual associations evoked in reading the primary objects of reflection (Rosenblatt, 1938, 1995). *Envisionment* entails four interwoven ways of looking at a text: (a) as a structural whole, (b) as a space to be explored, (c) as an object to be evaluated, and (d) as an opportunity for personal reflection (Langer, 1995a, 1995b).

A Taxonomy of Literate Engagement with Educational Software

Student-software interactions resemble engaged literate activity. Indeed, when students "make sense" of educational software, they may employ the same interpretive skills needed in paper-based tasks. This is partly due to heavy reliance on alphanumeric symbols in both computer- and paper-based texts. But even when computer "texts" use symbol systems not found in books (e.g., audio presentations, animation, video, etc.) and require kinds of interaction not possible with paper, these electronic texts require attention to the structure and organization of information, deployment of skills for locating and interpreting the information, and thoughtful meaning-making strategies. These skills are typical of engagement, self-regulated learning, problem-solving, and high literacy.

Bangert-Drowns and Pyke (2001) integrated notions of high literacy in a construct they called "literate thinking." Literate thinking entails abilities to evaluate the structure and content of texts (i.e., information representations) of various kinds, interpret texts from various perspectives, and reflect on issues of personal meaning in light of texts. They searched for instances of literate thinking with electronic texts in naturalistic observations of elementary school children interacting with conventional educational software. Bangert-Drowns and Pyke iden-

Table 1 □ Modes of student engagement with educational software

Name of mode	*Description*
Three problematic forms of engagement:	
Disengagement	Student avoids or discontinues software interaction; sometimes inattentive, purposeless, disinterested tinkering with software elements.
Unsystematic engagement	Student shows no higher-order goals with software; moves from one activity to another without apparent reason.
Frustrated engagement	Student attempts to achieve specific software goals unsuccessfully.
Competent engagement:	
Structure-dependent engagement	Student navigates and operates the software competently to pursue goals communicated by the software or teacher.
Three increasingly personalized and sophisticated forms of engagement:	
Self-regulated interest	Student adjusts software features to sustain deeply involved, interesting, or challenging interactions for personally defined purposes.
Critical engagement	Student manipulates software to test personal understanding or operational or content-related limitations of software representations.
Literate thinking	Student explores software from multiple, personally meaningful perspectives; uses perspective-sensitive interpretations to reflect on personal values or experiences.

tified seven distinct modes of student engagement with educational software that could be arranged hierarchically according to the degree to which they involved strategic and complex interpretive acts approximating literate thinking. (See Table 1.) The authors achieved 76% agreement in using the seven-level taxonomy to code the initial observation transcripts. In 53 subsequent observations (a total of 78 observations of 43 students), the authors found "the taxonomy proved a very robust means for describing student transactions with software irrespective of student characteristics or software types" (p. 219).

The three simplest forms of student engagement with educational software, (a) disengagement, (b) unsystematic engagement, and (c) frustrated engagement, were considered problematic. *Disengagement,* avoidance of software interactions, precluded meaningful involvement. In *unsystematic engagement,* students activated software features without apparent goal or coordination. Students attended to the surface features of the software rather than its content. *Frustrated engagement* described cases in which students pursued goals for the software interaction but were unable to realize them because of some lack of navigational, operational, or content knowledge. These levels of engagement resemble characteristics of student clusters that Ainley (1993) labeled "detached," "disengaged," "hopeful," and "keen-to-do-well."

Structure-dependent engagement, the fourth level of the engagement taxonomy, described software interactions where students competently navigated and operated the software, working in accordance with the directions of the teacher or the software. This form of engagement most resembled the characteristics of Ainley's (1993) "engaged" cluster. For Ainley, only students in the "committed" cluster evidenced high involvement, high ability, and use of "deep" learning strategies. The taxonomy, however, articulates three distinct forms of high-functioning engagement. In *self-regulated interest,* students used well-developed navigational and operational competence to pursue software features of personal interest. In *critical engagement,* students systematically explored the nature of the software content and its representations to test their capacities, adequacy, and validity. This form of engagement most resembles full-blown self-regulated learning. In *literate thinking,* students interpret software content from multiple perspectives with particular attention to the personal significance of the software experience.

Thus the Bangert-Drowns and Pyke taxonomy of literate engagement extends Ainley's cluster analysis by giving further articulation to the highest forms of engagement and by suggesting a hierarchical relationship among engagement forms. Bangert-Drowns and Pyke (2001) found this arrangement logically appealing, consistent with theoretical conceptualizations of self-regulated learning (e.g., Butler & Winne, 1995), and consistent with the engagement coding scheme developed by Lee and Anderson (1993). Lee and Anderson's lowest engagement levels (5–7) describe behavioral and cognitive disengagement, their behavioral engagement (level 4) describes unsystematic engagement, their behavioral and ambiguous cognitive engagement (level 3) describes frustrated engagement, their cognitive and behavioral engagement (level 2) resembles structure-dependent engagement, and their self-initiated cognitive and behavioral engagement corresponds to the highest three levels of the Bangert-Drowns and Pyke taxonomy.

Computer interactions are not all equal; some reflect more sophisticated interpretive activity than others and are likely related to enhanced learning. Teachers who use computers in their instructional activity would do well to monitor student engagement, encouraging students in problematic forms of engagement to higher levels. Teachers could use assistance in this effort; in 1998, only 20% felt "very well prepared" to integrate technology in their teaching (U.S. Department of Education, 1998). Does a taxonomy of literate engagement make sense to teachers in classroom practice, and can teachers distinguish among these forms of engagement?

Teacher Ratings of Student Achievement and Motivation

No prior research was found on classroom teacher ratings of student-learning engagement as a multidimensional construct. However, there are studies of teacher judgments of student achievement and student motivation. Because engagement entails cognitive and motivational components, these literature sources might hint at teacher competence for judging engagement.

In general, teachers assess student academic performance adeptly. In the Early Childhood Longitudinal Study (ECLS) (Perry & Meisels, 1996), following 23,000 children from kindergarten through the fourth grade, teachers made accurate assessments of student academic performance, sometimes predicting future achievement better than standardized measures. Similarly, Wright and Wiese (1988) found high correlations (ranging from +.57 to +.71) between teacher ratings of achievement and student performance on standardized achievement tests.

Hoge and Coladarci (1989) reviewed 17 studies yielding 56 measures of the concurrent validity of teacher judgments of student achievement. Correlations between achievement test performance and teacher judgments in the form of ratings (*Mdn* = +.61) were generally lower than rankings, grade equivalent judgments, estimated number of test items correct, and item-by-item judgments of test performance (*Mdn* correlations from +.67 to +.76). The authors concluded that a teacher could better judge a student's item-by-item test performance than give a general rating of a student's achievement across multiple tasks.

Of course, not all teachers were equally adept at rating student learning. One study (Helmke & Schrader,1987) reviewed by Hoge and Coladarci (1989), for example, reported teacher rating–achievement score correlations ranging from +.03 to +.90 among 31 teachers. Also, when asked to rate achievement on a common scale, teachers showed the same achievement rankings of students,but calibrated scales differently, showing positive or negative biases in their ratings.

Learning engagement has not only a cognitive component, but a motivational one as well. Are teachers effective in assessing learner motivation? Wright and Wiese (1988) found that teachers rated student effort and achievement on distinguishable, correlated criteria. These authors found no motivation criterion to validate the teachers' effort ratings and did not investigate the relations between effort ratings and standardized test achievement.

Sweet, Guthrie, and Ng (1998) asked teachers to rate students' "motivation to read" on six criteria developed from theoretical literature on intrinsic motivation and in collaboration with teacher focus groups. Teacher ratings on these six criteria were significantly and positively correlated with reading report card grades. Skinner and Belmont (1993) asked teachers and students to rate student behavioral and emotional engagement (a motivational construct) in both semesters of an academic year. Behavioral engagement entailed student effort, attention, and persistence during the initiation and execution of learning activities. Emotional engagement included interest (vs. boredom), happiness (vs. sadness), anxiety, and anger. Teachers and students were highly consistent in their ratings across semesters, with higher correlations for behavioral engagement ($r = +.72$) than for emotional engagement ($r = +.60$). Teacher ratings of student behavioral engagement were positively correlated with student ratings (about +.33 across semesters). However, teacher and student ratings of emotional engagement showed much less correspondence (+.21 and +.08 in fall and spring semesters, respectively). It is possible that ratings of emotional engagement were more highly inferential than behavior engagement and thus less reliable.

If different kinds of engagement differentially support learning, teachers should distinguish productive and counterproductive forms to adapt instruction. If teachers can accurately assess student engagement with educational software, their ratings could prove important moderators in research on learning with computers. The current study gathered qualitative and quantitative data to investigate issues regarding teacher rating of student engagement with educational software, with the ratings guided by descriptions developed from a seven-level taxonomy of literate engagement. The study sought to document the kinds of successes and obstacles that might typify teachers' attempts to rate student engagement in real classroom conditions. Research questions were:

- Is a seven-level taxonomy of engagement consistent with teacher experience of student software interactions?

- What difficulties do teachers encounter when rating student software engagement?
- Are teacher ratings of student software engagement meaningful? To what extent do teacher ratings of the software engagement of the same student agree? Do teacher ratings of software engagement correspond to other indicators of student literacy competence?

METHOD

Participants and Setting

The researchers collected data at an elementary "magnet" school for science and technology. Because of its magnet status, this school attracted students from across its urban school district, ensuring that its population possessed diverse racial, ethnic, socioeconomic, and academic backgrounds. With a technology emphasis, students and teachers regularly interacted with computers in the school, and frequently outside the school as well. The school employed many software types in its curriculum, including tools (such as word-processing, graphics, and database software), simulations, tutorials, games, and Internet browsers.

The researchers chose to work with the fifth grade, hoping that older students might demonstrate varied styles of software engagement. The school's two fifth-grade classroom teachers agreed to observe student computer interactions in their regular classrooms and during weekly periods in the school computer classroom. The computer classroom teacher also participated in the study. We were uncertain about whether the teachers would understand differentiations in modes of engagement and the particular forms that would guide their ratings. Given this uncertainty and the teachers' large investment of time, we invited participation from only these three teachers in this exploratory study.

Materials

Rating sheets were constructed for this study by paraphrasing essential characteristics for each mode of engagement from Bangert-Drowns and Pyke (2001; see appendix). The descriptions were numbered and arranged in order from disengagement through literate thinking. Teachers could indicate on a four-point scale (*almost always, often, rarely, never*) how frequently they had observed students interacting with software in each of seven different ways.

No engagement measure exists to compare to the teacher ratings of student engagement. However, because software engagement is a strategic, interpretive activity, it might correlate positively with reading test scores. We took the students' standardized fourth-grade reading test scores as a point of comparison (Degrees of Reading Power [DRP] Standard Test, Touchstone Applied Science Associates, 1997). The DRP tests are untimed reading comprehension tests administered as a normal part of this school's assessment program.

Procedure

The researchers met with the teachers to explain the rationale of the study, briefly review the seven forms of software engagement, discuss the rating forms, and invite questions for clarification. The teachers indicated that they understood the rating descriptions and thought it feasible to conduct the ratings. The researchers secured teachers' agreements to participate, and sought parental permissions and student informed assents to conduct the study. Researchers were authorized to obtain teacher ratings for 42 students and reading test scores for 31 students.

The two fifth-grade classroom teachers and the computer teacher received one rating sheet for each of their fifth-grade students. Teachers were asked to rate how frequently each student engaged in seven different kinds of interaction with software, based on direct observation. The classroom teachers each had 21 students to rate; the computer classroom teacher rated all 42 fifth-grade students. Teachers were asked to complete their ratings within a month. Standardized fourth-grade reading test scores were obtained for most of the rated students.

Rating conditions typified ordinary classroom situations where teachers have scant time for specialized training and must make numerous assessments of student motivation and learning on the basis of observation. Had measurement precision been the priority of this study, as in instrument development or quasi-experimental comparison, teacher training would have been a means to improve rater reliability. However, with formal training, we might have lost an opportunity to have teachers judge the correspondence of the taxonomy with their ordinary experience of student-computer interactions. We hoped that the teachers would use the rating forms as guides, but allowed them to operate in their classrooms in an authentic way. If the teachers could rate the students with little formal support from researchers, the rating rubric would more probably make meaningful contributions to ordinary classroom instruction. Potential hierarchical relations among engagement modes were not explicitly mentioned. The computer classroom teacher knew of the engagement taxonomy from reports of the researchers' earlier work. The two classroom teachers, however, knew little of the notions of engagement at the outset of their participation.

The three teachers were urged to base their judgments of engagement on direct observations of student computer interactions. Teachers were expected to summarize their impressions of each student working across different types of software with different tasks in both the regular and computer classroom. They were asked to complete the rating forms in a month's time.

After the rating data were collected and analyzed, a descriptive report was returned to the teachers. The researchers subsequently met with the three teachers as a group to discuss their evaluation of the taxonomy and their experience in its application. The group interview was unstructured but sought the teachers' evaluations of the taxonomy of engagement modes and the rating forms as well as their impressions of the rating process and results.

RESULTS

Teacher Evaluations of the Taxonomy of Student Engagement

In the postrating interview, the teachers reported that the taxonomy of student engagement was very descriptive of their students' work with educational software. They observed no instance of student-computer interaction that was not describable in terms of the taxonomy, and, conversely, they felt that they had observed instances of each of the taxonomy's levels.

Though the computer teacher had passing familiarity with the taxonomy of engagement from previous communications with the researchers, the two classroom teachers had not realized that the seven descriptions could be arranged in a meaningful order. When a sequential order among the seven modes of engagement was described to them, they agreed that the framework was sensible, made the individual levels more understandable, and implied new possibilities for instruction to enhance student engagement. All three teachers stated that the level descriptions and taxonomic framework helped them to interpret their students' interactions with the computer in new ways and focused their attention on specific features of student-computer interactions.

Teachers' Self-Reported Rating Activity

The teachers indicated the frequency of seven different forms of engagement for each student. Not surprisingly, the teachers found this time-consuming. The teachers observed that most students displayed a consistent pattern of software interaction, but they noted that engagement could fluctuate depending on the software and the classroom context. They felt unsure about how many observations were needed to develop stable impressions of engagement. As directed, all three teachers worked independently and made specific observations of students during the rating period, but they admitted extrapolating from their observations on the basis of other knowledge of the students to offset the perceived context sensitivity of engagement.

The teachers expressed some difficulties with the student engagement rating form. The extremes of the frequency ratings, *almost always* and *never,* seemed too extreme; Teacher A avoided using those labels. They found some engagement descriptions ambiguous, and some levels seem to employ multiple criteria. For example, the two classroom teachers were reluctant to identify instances of literate thinking because the rating form seemed to require students to explore multiple interpretations of their software experience *and* to communicate verbally with teachers about their software interactions. They felt that their direct conversations about specific software were too few to warrant a positive rating at this level. In general, the teachers thought disengagement, unsystematic engagement, and frustrated engagement (the "lower" levels of the taxonomy) were easier to rate because their behavioral indicators were most obvious.

The teachers felt confident of the majority of their student ratings. On a four-point scale from *very unsure* to *very sure*, Teacher A rated herself as sure on 18 of 21 ratings (86%); Teacher B rated herself as sure or very sure on 17 of 21 ratings (81%). The computer teacher indicated greater confidence in her ratings, marking herself as sure or very sure on 41 of 42 ratings.

The teachers stated that they made their engagement ratings independent of any knowledge of student reading competence. Indeed, though they found the taxonomy a useful framework, they resisted the notion that student-software interactions might be akin to literate activity. The teachers' primary goals in classroom computer use were to enhance competent computer use and to augment normal academic tasks. They volunteered examples where students who were competent readers did poorly while working on the computer, and vice versa.

Teacher Ratings of Student Engagement and Their Correlates

Table 2 presents the average ratings of each teacher for each form of engagement. All three teachers give their highest average frequency rating to structure-dependent engagement. Also, all three teachers reported "dysfunctional" forms of engagement (disengagement, unsystematic engagement, and frustrated engagement) less frequently than "functional" forms.

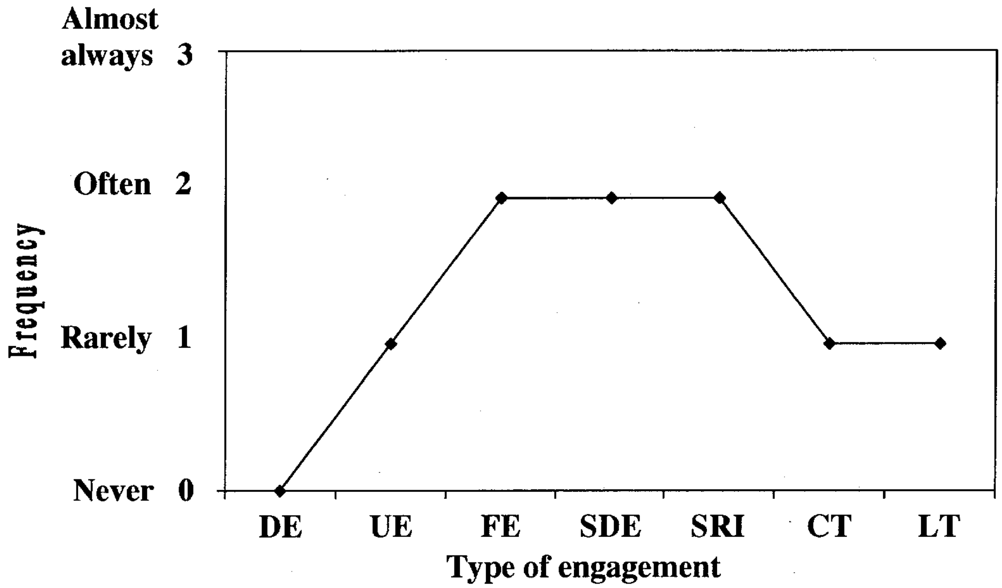
The teachers' ratings differed in important ways. For example, the computer teacher reported higher frequency of literate thinking than did the classroom teachers. The average rating for literate thinking by the computer teacher was between *often* and *almost always*. Both classroom teachers, on average, rated literate thinking as less than *rarely*. In addition, Teacher A's ratings tended to be more moderate than those of her colleagues, ranging from .67 to 1.62. (Recall that Teacher A had said she avoided using the *never* and *almost always* ratings. Teacher B's average ratings ranged from .10 to 2.76; the computer teacher's average ratings ranged from .38 to 2.38.) Teacher A also seemed

Table 2 □ Means and standard deviations for student engagement ratings given by two fifth grade teachers and the computer classroom teacher.

	Class A				*Class B*			
	Teacher A		*Computer Teacher*		*Teacher B*		*Computer Teacher*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Literate thinking	0.91	(0.70)	2.38	(0.67)	0.10	(0.44)	2.24	(0.70)
Critical engagement	1.14	(0.66)	2.43	(0.60)	2.05	(0.67)	2.29	(0.64)
Self-regulated interest	1.24	(0.63)	2.52	(0.68)	2.19	(0.68)	2.29	(0.72)
Structure dependence	1.62	(0.50)	2.71	(0.56)	2.76	(0.44)	2.38	(0.59)
Frustrated engagement	0.71	(0.64)	0.19	(0.51)	0.52	(0.60)	0.57	(0.75)
Unsystematic engagement	1.14	(0.66)	0.29	(0.64)	0.48	(0.68)	0.57	(0.75)
Disengagement	0.67	(0.58)	0.24	(0.54)	0.43	(0.51)	0.38	(0.67)

Note: Ratings were made on a 4-point scale (0 = *never*, 1 = *rarely*, 2 = *often*, 3 = *almost always*). In each class, *n* = 21.

Figure 1 □ Teacher ratings of engagement frequency for one student on each of seven types of engagement. (See Table 1 for types of engagement.)



to be a more conservative rater, giving, on average, lower frequency ratings to higher levels of the taxonomy, and vice versa, than the other teachers.

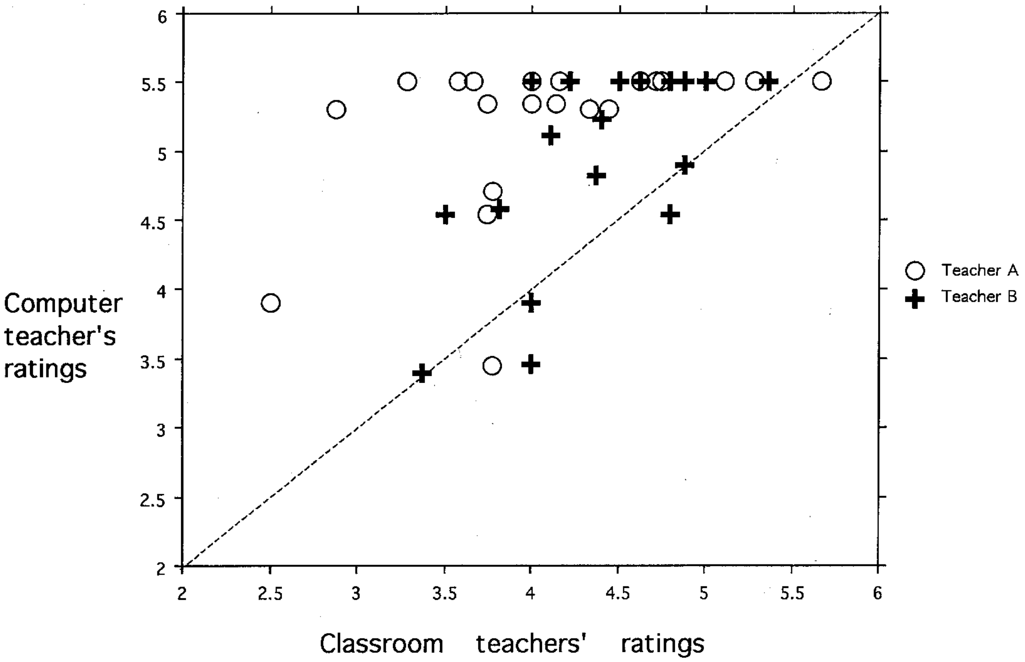
Because each student's seven ratings assessed very different behaviors, a sum or simple average of each student's ratings or a measure of internal consistency would not be meaningful. However, all three teachers rated the seven levels of engagement as if along a continuum, identifying a single level, or a set of contiguous levels, as most prevalent for most students. For example, Figure 1 shows seven engagement ratings given by Teacher A for one of her students. Three contiguous forms of engagement (frustrated engagement, structure dependence, and self-regulated interest) were given ratings of highest frequency. The computer teacher gave a *most prevalent* rating to 40 of 41 students (98%); the classroom teachers to 36 of 41 students (88%).

Because teacher ratings indicated predominant levels of engagement as if along a continuum, we summarized each student's seven ratings in single, frequency-weighted average ratings. Given values 1 through 7 to disengagement through literate thinking, one can multiply each engagement level by its teacher-rated frequency (0 to 3) and sum these products. This sum is divided by the sum of the teacher's frequency ratings to derive a single frequency-weighted engagement score. The seven ratings in Figure 1, for example, transformed to a single score of 4.33; the student's frequency-weighted rating is a bit higher than structure-dependent engagement.

Frequency-weighted average ratings could vary from 1 (student is always disengaged) to 7 (student is always engaged in literate thinking). In fact, these students' ratings varied between 2.5 and 5.7. The teachers rated their students positively; only 5 of the computer teacher's ratings (12%) and 12 of the classroom teachers' ratings (26%) were below structure dependence.

Figure 2 compares the engagement ratings of the computer teacher to those of the classroom teachers. If there were perfect agreement between the classroom teachers and the computer teacher in their ratings, the scores would fall along the diagonal. Instead, the scores tend to bunch above the diagonal, indicating that the computer teacher was routinely more positive in her ratings of student engagement. The computer teacher's averaged ratings (*M* = 5.09) were

Figure 2 □ Scattergram comparing two classroom teachers' ratings of student engagement to a computer teacher's ratings.



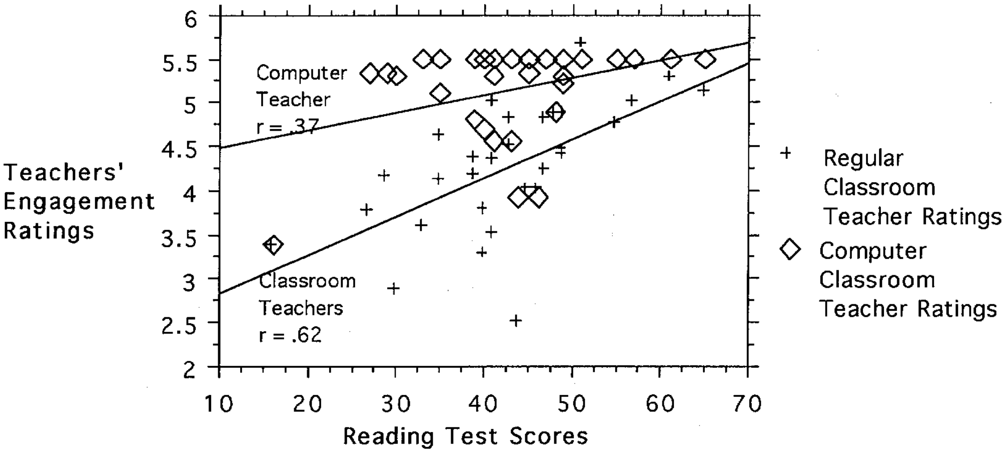
significantly higher (Cohen's d = +1.23, paired $t(40) = 7.8$, $p < .0001$) than the two classroom teachers' combined (M = 4.28), though the two teachers gave similar average ratings to their own classes (M [Teacher A] = 4.10, M [Teacher B] = 4.46, Cohen's d = +.55, $t(40) = 1.73$, $p = .09$).

The teachers explained the differences in magnitude in two ways. One explanation was an artifact of the rating form: Teacher A reported that she was very unlikely to rate any student at the highest level of engagement because she interpreted the description of the level as requiring verbalization on the part of the students; she rarely discussed the software with the students. The second explanation related to differences in classroom contexts. Teacher A, the most conservative of the three raters, usually required students to accomplish specific tasks on the computers in her classroom. Given the structured quality of her assignments, students might not have as much freedom to explore areas of personal interest, as in the higher levels of engagement. Teacher B and the computer teacher permitted more open-ended explorations of software.

In addition, the computer teacher worked more intensively with educational computing, during and after the school day. The classroom teachers believed that the computer teacher knew more about the students' computer interactions and could discuss higher levels of engagement that were difficult to observe. The computer teacher reported herself as being more confident than the other teachers in her ratings. On a four-point scale from *very unsure* (rating of 1) to *very confident* (rating of 4), she rated herself at 3.31 (greater than *confident*) while the classroom teachers each rated themselves exactly 2.86 (less than *confident*). The difference between the confidence ratings of the two classroom teachers and the computer teacher was statistically significant (F (1, 82) = 15.93, p = .0001, Cohen's d = +.85).

In spite of differences in rating magnitude, the teachers agreed significantly about the rank assigned to the students' engagement. The correlation between the computer teacher ratings and the two classroom teacher ratings was r = +.47 (F (1, 40) = 11.61, p = .002). The Spearman rank correlation corrected for ties was r_s = +.52, p

Figure 3 □ Scattergram showing regression lines linking teacher ratings of student engagement with student reading test scores.



= .0008, n = 41. (The uncorrected r_s was +.56.) However, the correlation was stronger between Teacher B and the computer teacher (r_s = +.68, p = .002, n = 21) than between Teacher A and the computer teacher (r_s = +.44, p = .05, n = 21).

Teacher ratings of student engagement showed a significant correlation with students' standardized reading test scores (Figure 3). For the two classroom teachers combined, this correlation was r = +.62, p = .0002, n = 31. The ratings of Teachers A and B separately correlated with reading scores at r = +.62 (n = 17) and r = +.68 (n = 14), respectively. When asked during the postrating interview about the high correlations with reading scores, the teachers insisted that they consciously put aside their knowledge of the students' achievement in classroom activities and standardized tests. Indeed, the teachers felt certain that reading ability and software interaction were independent and that some of their students were more deeply engaged with software than they would have judged from their performance on other academic tasks.

The computer teacher's ratings showed a lower, but significant, correlation with student standardized reading test scores (r = +.37, p = .04, n = 31). However, the correlation was much higher for one fifth-grade class than the other. The correlation with the reading scores of Teacher B's students was +.67 (p = .009), a magnitude consistent with the correlations observed with the two classroom teachers' ratings. The computer teacher's ratings for students of Teacher A only correlated +.10 with reading scores. During the postrating interview, the computer teacher reported rating the students of one class before rating those of the other, but she could not recall which was done first, nor whether she changed rating strategies between classes.

The computer teacher rated engagement differently from the regular classroom teachers. In 38 of her 42 ratings (90%), she treated engagement dichotomously. Most of her ratings reported low frequencies of the three problematic levels of engagement and high frequencies of the highest four levels of engagement. In a few cases, the dichotomy was reversed; a student was rated as unlikely to manifest high engagement levels, but likely to manifest the lowest levels. These dichotomous ratings were restricted in range to the positive end of the scale. Two thirds of her 42 ratings were 5.3 or above (i.e., at self-regulated interest or above). This positive bias in engagement ratings was somewhat more pronounced in her ratings of the students of Teacher A than in those of Teacher B. Restriction of range and nonnormality may have contributed to the small correlation with reading scores of Teacher A's students.

DISCUSSION

Learning engagement, cognitive, and affective involvement of students in learning tasks, can enhance learning achievement. Learning engagement has been studied unidimensionally, as depth-of-processing, time-on-task, or intrinsic motivation (e.g., Jacques, Preece, & Carey, 1995; Kumar, 1991; Martens et al., 1997; Skinner & Belmont, 1993). However, such notions can mask engagement's complexity. Two students may be engaged to the same degree, but in very different ways, and different forms of engagement may predict different kinds of learning. Some researchers studied engagement multidimensionally (e.g., Ainley, 1993; Bangert-Drowns & Pyke, 2001; Corno & Mandinach, 1983; Lee & Anderson, 1993; Nystrand & Gamoran, 1991).

Students enthusiastically engage in computer interactions, but such enthusiasm does not always translate into meaningful learning. Bangert-Drowns and Pyke (2001) defined a taxonomy of seven forms of engagement that can shed light on student learning with educational software. First, viewing learning with software as a literate act, a view echoed in other research (e.g., Jacobson & Spiro, 1995; Scardamalia, et al., 1989; Yang, 2002), allows educators to apply literacy research to pedagogical problems in technology integration. Second, the taxonomy integrates theoretical constructs (e.g., disengagement, software navigation, self-efficacy, intrinsic motivation, self-regulation, critical thinking, and high literacy) into a coherent, hierarchical framework. Third, it views student engagement, not as a dichotomous quality (engaged-disengaged), but as a multidimensional one; not all engagement is equal in its qualities and effects on learning.

Though theoretical aspects of the engagement taxonomy require further clarification (Bangert-Drowns and Pyke, 2001), this study considered measurement issues. Can forms of engagement be identified by teachers in authentic classroom situations? Teachers feel ill-equipped to integrate computers in their regular curriculum (U.S. Department of Education, 1998); a taxonomy of engagement might help initiate, guide, and assess efficacious student-computer interactions. Teachers are fairly adept at rating student achievement and motivation (e.g., Hoge & Coladarci, 1989; Perry & Meisels, 1996; Skinner and Belmont, 1993). However, no research has investigated teacher assessments of engagement, particularly with computers, in authentic classroom conditions.

Teacher ratings of student engagement in this study resembled ratings for student motivation and achievement obtained in other research. Hoge and Coladarci (1989) found that teacher ability to judge student achievement, though generally good, varied considerably. Similarly, in this study, teachers rated engagement frequencies differently when assessing the same students. Differences in magnitude are due in part to real differences in teacher knowledge of their students' computer experiences. However, differences are also likely to result from positive and negative rater biases.

In spite of differences in rating magnitude, teachers substantially agreed in ordering students by level of engagement (r_s = +.52). Furthermore, engagement ratings of the two classroom teachers were significantly correlated (*r* = +.62) with student fourth-grade standardized reading scores. For Teacher B's students, the computer teacher's engagement ratings had a similarly large correlation with reading score (*r* = +.67), but for Teacher A's students, the correlation was small (*r* = +.10). Such variation in correlations also was found in studies reviewed by Hoge and Coladarci (1989).

Several factors made the rating task difficult for these teachers. First, rating the frequency of seven different forms of engagement across multiple context for each student was time consuming. Second, at least one teacher reported avoiding using the end points of the rating scales (*never* and *almost always*) because they seemed too extreme. Third, engagement descriptions were differentially difficult to assess. Some descriptions seemed to suggest multiple conditions that had to be met conjointly; other descriptions seemed more highly inferential or required student interview.

Despite the study's ecological validity, several threats to internal validity exist. The three teachers constitute a small sample of unknown generalizability. Difficulties with rating forms contributed an unknown amount of "noise" to teacher assessments of student engagement. Raters were not trained in order to enable the researchers to explore teachers' engagement ratings with minimal preparation. Raters may not have completed their tasks as reliably as desired. Finally, the raters were not blind to personal qualities of rated students. This typifies assessments in authentic instructional situations, but such personal knowledge could bias teacher ratings in various ways. An exploratory design that emphasized authenticity seemed most appropriate for this first attempt to investigate the capabilities of teachers to rate student engagement with computers multidimensionally in real classrooms. Replications with different degrees and kinds of control of extraneous variables are warranted to lend greater confidence to this study's findings.

This study, as others on teacher judgments of achievement and motivation, suggests that teachers rate more reliably across groups of students and when students are ranked. Engagement ratings of individual students differ across raters because of the complexity of the phenomenon, the different ways in which students can be known, and rater biases. Some strategies could increase the rating reliability. Training for raters, and concrete and specific descriptions of engagement would help. Less time-consuming rating tasks (e.g., providing a single engagement rating along a continuum rather than seven separate ratings) would reduce the risk of rater fatigue. Hoge and Coladarci (1989) found that teachers could better judge students' item-by-item test performance than predict achievement globally. Teachers might show greater agreement in engagement ratings when asked to assess specific instances of student-software interaction than make more global judgments.

How precise must a teacher's rating of student engagement be in actual classroom practice? Instructionally, the best engagement ratings would be precise, individualized, and situational. If a teacher knew how a student was engaged at any given moment with a particular learning task, the teacher could adapt instruction situationally. For a student momentarily unsystematically engaged, the taxonomy of engagement might suggest support for fashioning and persisting in longer-term goals to advance to structure dependence. For the student in self-regulated interest, the taxonomy might suggest stimulation to evaluate knowledge and knowledge representation in order to progress to critical engagement. Teachers often judge students' momentary motivations and comprehension; such situationally specific assessments of engagement are entirely possible, but best suited to tutorial and individualized conditions. As long as schools require teachers to teach to classes, situational assessment of all students' engagement at all times will be largely impossible.

Alternatively, teachers in this study identified styles of software interactions across different software, learning tasks, and physical and interpersonal contexts. If engagement predispositions are trans-situational and relatively enduring, teachers could make longer-term curricular plans to present information in ways consistent with styles of different groups of students or to enhance engagement for all. Under these circumstances, precision of teacher engagement ratings may be much less important. To know which students are capable of higher engagement, which hover around structure dependence, and which are often stuck in dysfunctional forms of engagement may be sufficient for curricular planning and for monitoring the general progress of the class.

Though the Bangert-Drowns and Pyke taxonomy of engagement was not the focus of this research per se, the study has implications for the taxonomy. The three participating teachers found the different forms of engagement recognizable in and sufficient to describe student computer interactions, even without formal training. When told of hierarchical relations among engagement levels, the teachers better understood the nature of the students' work, and ways to enhance it. In Skinner and Belmont (1993), teacher and student ratings of aspects of student engagement correlated across two semesters. The teachers in this study also found consistency in student engagement, identifying predispositions for forms of engagement across software situations.

Finally, teacher ratings of student software engagement correlated significantly with student fourth-grade reading scores. This suggests correspondence between literate engagement with educational software and literacy as measured in a conventional, paper-based test. It is possible that teacher awareness of student reading abilities influenced their judgments of engagement. However, several factors operate against that possibility. First, reading tests were completed in fourth-grade classes more than a year before teachers made their ratings for this study. Second, the teachers did not think that software interactions were literate acts. Third, the computer teacher had little to do with assessing student academic work. Fourth, none of the teachers had ever evaluated student computer use as an academic activity in itself. The teachers' chief interest in computers was as a tool to augment other academic activities, not as an object of literacy. Finally, teachers made seven separate ratings of each student's engagement, and the teachers did not seem aware that the seven ratings could be related hierarchically.

In spite of no training, a difficult rating form, and a moderately inferential rating task, these teachers found considerable usefulness and agreement in their efforts to evaluate the qualities of their students' engagement with multimedia. Such agreement in this exploratory study suggests promising classroom applicability for the seven-level taxonomy of student engagement. □

Robert L. Bangert-Drowns [rbangert@csc.albany.edu] is Associate Professor at the University at Albany, State University of New York.

Curtis Pyke is Assistant Professor at The George Washington University, Graduate School of Education and Human Development, Washington, DC.

This research was conducted under the auspices of the National Research Center on English Learning and Achievement (CELA), operated by the University at Albany in collaboration with the University of Wisconsin-Madison. The Center is supported by the Office of Educational Research and Improvement, U.S. Department of Education (Award # R305A60005). However, the views expressed herein are those of the authors and do not necessarily represent the views of the department.

REFERENCES

Ainley, M.D. (1993). Styles of engagement with learning: Multidimensional assessment of their relationship with strategy use and school achievement. *Journal of Educational Psychology, 85*(3), 395–405.

Bangert-Drowns, R.L., & Pyke, C. (2001). A taxonomy of student engagement with educational software: An exploration of literate thinking with electronic text. *Journal of Educational Computing Research, 24*(3), 213–234.

Bereiter, C., & Scardamalia, M. (1987). An attainable version of high literacy: Approaches to teaching higher-order skills in reading and writing. *Curriculum Inquiry, 17*(1), 9–30.

Butler, D., & Winne, P. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281.

Corno, L., & Mandinach, E.B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist, 18,* 88–109.

Guthrie, J.T. (1996). Educational contexts for engagement in literacy. *The Reading Teacher, 49*(6), 432–445.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education, 3,* 91–98.

Hoge, R.D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297–313.

Jacques, R., Preece, J., & Carey, T. (1995). Engagement as a design concept for multimedia. *Canadian Journal of Educational Communication, 24*(1), 49–59.

Jacobson, M.J., & Spiro, R.J. (1995). Hypertext learning environments, cognitive flexibility, and the transfer of complex knowledge: An empirical investigation. *Journal of Educational Computing Research, 12*(4), 301–333.

Kearsley, G., & Shneiderman, B. (1998). Engagement theory: A framework for technology-based teaching and learning. *Educational Technology, 38*(5), 20–23.

Kuh, G.D. (2000). *The NSSE 2000 report: National benchmarks of effective educational practice.* Bloomington, Indiana: Center for Postsecondary Research and Planning, Indiana University.

Kumar, D.D. (1991). A meta-analysis of the relationship between science instruction and student engagement. *Educational Review, 43*(1), 49–61.

Langer, J. (1995a). Literature and learning to think. *Journal of Curriculum and Supervision, 10*(3), 207–226.

Langer, J. (1995b). *Envisioning literature: Literary understanding and literature instruction.* New York: Teachers College Press.

Lee, O., & Anderson, C.W. (1993). Task engagement and conceptual change in middle school science classrooms. *American Educational Research Journal, 30*(3), 585–610.

Martens, B.K., Bradley, T.A., & Eckert, T.L. (1997). Effects of reinforcement history and instructions on the persistence of student engagement. *Journal of Applied Behavior Analysis, 30*(3), 569–572.

Nystrand, M., & Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English, 25*(3), 261–290.

Perry, N.E., & Meisels, S.J. (1996). *How accurate are teacher judgments of students' academic performance?* Chicago, IL: National Opinion Research Center. (ERIC Document Reproduction Service No. ED418154)

Rosenblatt, L.M. (1938). *Literature as Exploration.* New York: Appleton Century.

Rosenblatt, L.M. (1995). Continuing the Conversation: A Clarification, *Research in the Teaching of English, 29*(3), 349–354.

Ryan, A.M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal, 38*(2), 437–460.

Scardamalia, M., Bereiter, C., McLean, R.S., Swallow, J., & Woodruff, E.. (1989). Computer Supported Intentional Learning Environments. *Journal of Educational Computing Research, 5*(1), 51–68.

Skinner, E.A., & Belmont, M.J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581.

Spiro, R.J., & Jehng, J.C. (1990). Cognitive flexibility and hypertext: Theory and technology for the nonlinear and multidimensional traversal of complex subject matter. In D. Nix & R. Spiro (Eds.). *Cognition, education, and multimedia: Exploring ideas in high technology* (pp. 163–205). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sweet, A.P., Guthrie, J.T., & Ng, M.M. (1998). Teacher perceptions and student reading motivation. *Journal of Educational Psychology, 90*(2), 210–223.

Touchstone Applied Science Associates (1997). *Degrees of Reading Power Standard Test.* Brewster, NY: Touchstone Applied Science Associates.

U.S. Department of Education. (1998). *Teacher Survey on Professional Development and Training, FRSS 65.* National Center for Education Statistics, Fast Response Survey System.

Wright, D., & Wiese, M.J. (1988). Teacher judgment in student evaluation: A comparison of grading methods. *Journal of Educational Research, 82*(1), 10–14.

Yang, S.C. (2002). Multidimensional taxonomy of learners cognitive processing in discourse synthesis with hypermedia. *Computers in Human Behavior, 18*(1), 37–68.

See Appendix A, next page.

Appendix A □ Teachers rated each student on seven types of software engagement using this form

Student ID ___________________________________

For each of the seven statements below, indicate how often you've seen the student engaged in the identified behavior.

1. Student stops interacting with the software. Student may sit and tinker with the software in a seemingly purposeless or disinterested way with little or no response to feedback from the computer. Or, student may in fact turn away from the software or resist using it at all.

Almost always Often Rarely Never

2. Student moves from one incomplete activity to another without apparent reason. Student successfully completes simple tasks within the software but does not link tasks for higher-order goals.

Almost always Often Rarely Never

3. Student tries to effectively interact with the software, but unsuccessfully. Student might manifest frustration in negative comments, confusion, aggressive behavior, erratic behavior, or signs of agitation, distress, or anxiety.

Almost always Often Rarely Never

4. Student pursues goals communicated by the software. Student may not yet display full mastery of software features, but responds to operational, navigational, or content organization.

Almost always Often Rarely Never

5. Student stimulates and maintains deeply involved interactions with the software. Student adjusts software features to sustain interesting or challenging interactions and creatively uses software for personally defined purposes.

Almost always Often Rarely Never

6. Student manipulates software features, keenly observes the effects of the manipulations, and integrates the results in future interactions. These manipulations seem designed to test personal understanding of the software content or the limitations of the software presentations.

Almost always Often Rarely Never

7. Student explores and develops multiple interpretations of a software experience. Student manipulates software features to explore different perspectives. In verbal statements, student describes different perspectives and use of software interactions as an opportunity to reflect on personal values or experience.

Almost always Often Rarely Never

>>How confident are you of your ratings above?

Very confident Confident Unsure Very unsure