Student Consistency and Implications for Feedback in Online Assessment Systems

Tara M. Madhyastha[1] and Steven Tanimoto[2]
madhyt@u.washington.edu, tanimoto@cs.washington.edu
[1]Department of Psychology, University of Washington
[2]Department of Computer Science, University of Washington

Abstract. Most of the emphasis on mining online assessment logs has been to identify content-specific errors. However, the pattern of general "consistency" is domain independent, strongly related to performance, and can itself be a target of educational data mining. We demonstrate that simple consistency indicators are related to student outcomes, and suggest how consistency might be used in an online assessment framework to provide scaffolding to help students in need.

1 Introduction

Online assessment systems have the potential to supply detailed information about how students interact with them, which can be used to provide useful feedback. Much of the effort in mining this data has focused on identifying student misconceptions and partial understandings, in an effort to build upon their existing knowledge and direct them towards corrective interventions. However, there is a more general type of information available from such systems that may be valuable to assess, which we call student consistency. Specifically, we refer to the ability of a student to self-appraise his or her performance while, in our context, interacting with a computer to complete some untimed task or assessment. For example, a student is inconsistent when he or she executes a command, obtains a result that differs from what students are told to expect, and takes no further action to resolve the problem. The evidence indicates that the student made a mistake, but the student does not acknowledge this through his or her actions. Because markers of consistency do not necessarily require pedagogical content models, they may be easier to define than sequences of logged actions that describe a certain misconception. They be used to give students helpful feedback or scaffolding, without detailed pedagogical content models.

This paper outlines a set of consistency markers developed in the context of a laboratory exercise conducted in a college-level course: Artificial Intelligence for Nonmajors. We demonstrate that consistency is related to outcomes, both specifically to the score obtained on the laboratory exercise, and more generally to performance on the final exam. We suggest how consistency markers might be used to provide formative feedback to the student.

2 Background

2.1 Consistency

Cognitive consistency was an important area of research in the 1950's and 1960's that sought to understand how people would behave as a function of the dissonance caused

when an individual simultaneously held two conflicting cognitions (e.g. beliefs, opinions). Theories posited that the conflicting cognitions would cause a quantifiable tension that the individual would seek to resolve in some way, recreating a consistent view. Unfortunately, consistency theories failed to fully explain behavior – individuals varied greatly in their ability to tolerate dissonance and their behavior was usually highly situation dependent [1].

Nevertheless, the idea of consistency is an important subtext in education. Teachers assume that student knowledge is consistent and that challenging their incorrect knowledge will cause students to attempt to reconcile the dissonance and learn. To this end, it is recommended that teachers should check the extent to which students hold erroneous concepts throughout instruction, ideally to deliver personalized feedback to students. This process is called "formative assessment" and feedback intended to help learners improve their performance is called "formative feedback". This approach usually requires a detailed model of the content domain and the ways in which students interact with it correctly and incorrectly. Constructing these models is time-consuming and difficult. Furthermore, there is little consensus on what constitutes appropriate formative feedback [2]. Complex, unspecific, or confusing feedback can have a negative effect on learning[3].

One reason feedback may be confusing may be that when students have not achieved mastery of a topic, their understanding is inconsistent. This was Sleeman's hypothesis for why addressing algebraic misconceptions was no more effective than reteaching (which is a simpler approach) [4]. Indeed, mathematical models that align specific errors to a linear ability scale find that groups of errors are "clustered", corresponding to a specific level of mastery [5]. Within each cluster of errors, students vary in the specific error they display. For example, a student at a relatively low level of mastery of graph reading may be equally likely at different times to interpret a straightforward graph of speed versus time with a constant positive slope as "the object is getting faster" (higher is faster) or "the object is getting slower" (higher is slowing down, as in up a hill).

Consistency itself is related to ability. This idea has been exploited to develop consensus-based assessments, where the correct response is defined as the one that most people agree on, and an individual score is measured as a distance from the consensus response. This approach has been used both to score people on situational judgment tests, and to measure intelligence [6]. When United States Military Academy students were asked to respond to a survey about their political beliefs (ranging from liberal to conservative) it was found that their consistency in these responses correlated with their SAT or ACT scores [7]. In research on physics misconceptions, we found that students who held any concept consistently, whether correct or incorrect, had higher math ability than students who were inconsistent [8]. The underlying logic for a relationship between ability and consistency is simple; a consistent response is one that many people can share by using similar underlying reasoning processes, but errors in reasoning and knowledge yield inconsistent, differing, responses. Furthermore, a consistent reasoning strategy requires a significant amount of knowledge about the topic and an ability to reconcile situations where one's beliefs do not match those of others (e.g., people who attempt to explain

other's dissenting beliefs by creating for themselves a clear explanation of the opposing view are likely to converge on consistent reasoning than those who dismiss alternative points of view).

Therefore, in this paper we conducted a study to examine how consistency, which requires no complex models of pedagogical content knowledge to measure, is domain-independent, and is easy to identify, is related to performance on online activities and assessments. We describe the implications that inconsistency has on the kind of feedback that might be appropriate in an online assessment system.

2.2 The PixelMath Software.

PixelMath is an educational image processing system. It was developed as a web-oriented successor to the "Pixel Calculator" program that, in turn, was developed by Tanimoto and associates as part of the NSF-funded project "Mathematics Experiences Through Image Processing" [8]. The purpose of PixelMath is to empower students to manipulate digital images using a mathematics-oriented, rather than an artist-oriented, interface. It provides special affordances that reveal the mathematical structure of each image and that can interpret mathematical formulas as image transformations and syntheses. PixelMath also provides a scripting facility using Python, making it possible to teach and learn programming in an authentic, "on-demand" context. PixelMath is both a tool for teaching and learning and a tool for educational research. It is hosted within an online learning environment called INFACT, which serves in part to collect activity data from students as they are learning using online tools. Whenever a student performs an operation in PixelMath, such as selecting a menu item, zooming in on the image, or running a mathematical formula to transform an image, a description of the operation is sent to the INFACT server, and it is stored in a database together with the user ID of the student and the time and date of the event. This makes it possible to perform formative assessment and/or data mining for the student activities without having the students take tests.

3 Method

We describe the research method used in this study.

3.1 Subjects

Subjects were students in one of the authors' Introduction to Artificial Intelligence (for nonmajors) course at the University of Washington in Winter quarter 2008. The University of Washington is highly selective, admitting approximately 100 students into the major each year, so there is a huge demand for courses for non majors from overlapping departments. We asked students to give consent to link INFACT log files from a laboratory exercise with course grades. Of the 40 students in the class, 34 completed the laboratory exercise. Thirty of those who completed the laboratory consented to participate in the study. Three did not submit the laboratory worksheet. Therefore, the final sample consisted of 27 students, including 2 females and 25 males of college age. This highly skewed gender distribution is typical among engineering classes.

The majority of these students were majors in the Applied and Computational Math Sciences program (N=17) or Electrical Engineering majors (N=5) and the remainder were divided among varied other non-computer science majors.

3.2 Procedure

Students were instructed to work individually, in a laboratory classroom with individual computers, to complete a series of exercises using PixelMath to understand some key concepts of image processing. These activities covered concepts of sampling, histograms, thresholding and morphology. They were allowed to ask questions of the instructor, which were answered individually, and to talk to each other. The lab was intended to take approximately an hour to complete. We observed little unrelated activity (e.g. web browsing or messaging) once students began working on the laboratory. The laboratory exercises included very specific instructions, including formulas to use to accomplish critical goals and to use as starting points for inquiry. Students were given participation points for turning in a completed worksheet. The laboratory session was held at noon on a Friday and students were given until Monday afternoon to turn in their work. Most students turned in their assignments by the end of the hour.

3.3 Data Logging and Coding

PixelMath logs a variety of timestamped activities, including file manipulations, image cloning, mouse clicks (zoom operations), and transformation formulas that students enter. From this data we extracted variables related to the amount of time students spent using PixelMath, errors made, and student consistency. The time variables were total time spent, number of logged events, and average time between events. Parse errors were totaled for each student to create a count of errors.

We identified seven Boolean consistency indicators that we categorized in three groups as follows:

- Matches worksheet (2 indicators)

At various points, students were asked to write in a response on the worksheet and use that value in a subsequent calculation, or to try to accomplish some task and then report some value that they found. For example, in the sampling activity, students are asked to determine the minimum number of pixels required, theoretically, to represent a long, white, picket fence. They are then asked to use this number of pixels to create a downsampled image. (Figure 1 is a screen shot that shows the original image in one window, the isolated picket fence in another, and the downsampled picket fence in a third; the PixelMath calculator is also visible.) Later, students are asked to report the sampling factor that resulted in the minimum number of pixels necessary to view the picket fence with minimum loss of pixels. We record a Boolean flag for each of the two worksheet responses: true if the worksheet matches a corresponding command (at any place in the log), and false if it does not. Note that “matching” does not indicate correctness; students often had matching consistency for incorrect answers in the log and worksheet.

- Logical consistency (1 indicator)

Students are given a formula to downsample an image and are asked to revise it so that they can view the image using the minimum number of pixels. Increasing the scaling factor reduces the number of pixels, and decreasing the scaling factor increases the number of pixels. Therefore, an ideal searching behavior would converge upon the ideal scaling factor (ideally obtained through calculation) by moving in subsequently smaller steps around the goal. Inconsistent behavior would be represented by repeatedly trying identical scaling factors and/or taking repeated steps in the incorrect direction. Repeated behavior is defined as 2 or more times. We code this sequence as true if the student is consistent and false if they are not, or if they do not attempt the exercise.

- Recognized expected outcome (5 indicators)

In several places, the student is instructed to execute a specific command, such as to open a specific file or to perform some image transformation. If the student does not execute this command correctly, the output will not be expected and subsequent instructions will not make any sense. This is a consistency error of failure to reconcile the differing information. Ideally, the student should recognize this inconsistency and re-execute the command, or ask a fellow student or the instructor for help and re-execute the command correctly. We coded failure to do so at any point in the lab as a false and a correct execution as true.

3.4 Outcome measures

We defined three outcome measures. The first was the grade on the worksheet. This was a score of 1-3 where 1 indicated an incomplete assignment (some questions were unanswered), 2 indicated some partial understanding of the key concepts, and 3 indicated a solid understanding of the concepts.

The second outcome measure was performance on questions related to image processing on the final exam for the course. The final exam consisted of a multiple choice section (Final Part 1) and an open ended section (Final Part 2) with three questions. The first question (Final Part2.Q1) was a direct analogue to the first activity involving sampling. The second question (Final Part2.Q2) was more difficult and required a deeper understanding of the concepts. The third question (Final Part2.Q3) was an in depth question on unrelated material. The final exam was administered 2 weeks after the laboratory session.

The third outcome measure was a metric reflecting general performance in the class. We obtained this measure by doing an unrotated principal components factor analysis on the subscores of all parts of the final. Two factors had eigenvalues over 1, resulting in a two factor solution that accounted for 70.37% of the variance in scores. The first factor loaded .89 on the score for Final Part2.Q1, .83 on the score for Final Part2.Q2, and .56 on the Final Part 1. The second factor loaded .96 on Final Part2.Q3. This pattern of loading suggests that the first factor represents general knowledge gained in the course, separate from the format of the exam (e.g., open-ended questions versus multiple choice). We use the first factor as an outcome measure of general class knowledge.

4 Results

Table 1 shows the mean and standard deviation for the outcome measures, measures of time spent completing the laboratory exercise, and the consistency measures. There was clear variation among the students on all measures, including the consistency measures that may seem obvious (for example, writing the same response on the worksheet that was used in the exercise). This variation is particularly substantial considering that these students are highly selected into a competitive state university, and are expected to have developed skills that would result in higher consistency measures than the population as a whole.

There is also significant variation in scores on the final exam. We note that the average score on Final Part2.Q2, the more difficult of the two questions dealing with image processing, is lower than the average score of Final Part2.Q1. This suggests that the questions were difficult enough to avoid ceiling effects.

Table 1. Summary statistics for outcome measures, general logfile measures, and consistency measures. Maximum score or range of scores is given, where appropriate, in parentheses following the measure.

Measure	Mean (SD)
Outcome measures	
Worksheet Score (1-3)	2.57 (.69)
Final Part 1 (50)	30.60 (7.01)
Final Part2.Q1 (10)	6.97 (3.64)
Final Part2.Q2(10)	5.87 (4.17)
Final Part2.Q3(10)	7.67 (2.63)
Other Logfile Measures	
Number log events	273.60 (262.87)
Average time between events (in minutes)	.21 (.08)
Number parse errors	2.13 (2.33)
Consistency measures	
Matches worksheet (2)	1.48 (.64)
Recognized expected outcome (5)	4.13 (1.66)
Logical consistency (1)	.37(.49)

To examine the differences in outcome measures based on consistency, we split the students by the median sum score of consistency (7) forming two groups. We call these the low consistency group (N=14) and the high consistency group (N=13). We conducted a one-way ANOVA to determine whether outcome measures differ across the two groups. Results are summarized in Table 2.

On average, the high consistency group scored higher on all outcome measures (though not all differences were significant). The high consistency group obtained significantly

greater worksheet scores (p=.002). The difference in scores is particularly noteworthy, because the consistency measures do not measure correct or incorrect responses, whereas the worksheet scoring does. Furthermore, the high consistency group scored marginally significantly higher on the Final Part2.Q1 (p=.083) and significantly higher on the Final Part2.Q2 (p=.049). These two questions of the final were designed to measure the same concepts covered by the worksheet, and were administered two weeks afterward. The high consistency group also scored significantly higher on the class knowledge measure extracted from the midterm scores (p=.018). The effect size (measured by Cohen's d) was quite large.

Differences in consistency cannot be attributed solely to "carelessness" or greater or lesser time spent on the assignment. The high consistency group logged fewer events (not significant) with more parse errors (marginally significant, p=.062). The average time spent between logged events was virtually identical among the two groups.

Table 2. Analysis of variance.

	High Consistency (N=13)		Low Consistency (N=14)				
	Mean	SD	Mean	SD	F	Sig.	Cohen's d
Worksheet Score	2.92	.28	2.21	.80	9.12	**.006**	**1.02**
Number Log Events/Student	198.31	82.55	326.71	357.81	1.59	.219	
Average Time Between Events (minutes)	.23	.08	.20	.09	.71	.409	
Number Parse Errors/Student	3.15	2.82	1.43	1.70	3.81	.062	
Final Part 1 (50)	33.23	6.09	29.43	7.00	2.26	.146	
Final Part2.Q1 (10)	8.46	2.26	6.14	4.07	3.27	.083	
Final Part2.Q2(10)	7.92	3.62	4.79	4.21	4.28	**.049**	**0.86**
Final Part2.Q3(10)	8.15	1.52	7.93	2.53	.08	.783	
Class Ability Measure (Factor extracted from Final Subtests)	.56	.58	-.28	1.04	6.46	**.018**	**0.84**

5 Implications for Feedback

We have proposed some rough indicators of consistency and shown that the low consistency students perform worse on several important class outcomes than the high consistency students. It is particularly remarkable that we find such variation in consistency in such a highly selected population.

We do not know the reasons why students are inconsistent, and we have not demonstrated whether higher consistency results in higher performance or whether it is a

side-effect of something else (e.g., low level of engagement or interest in the class). This is a topic for future research. However, the inconsistencies that we have coded are rather blatant. When a student repeatedly types in the wrong command and does not recognize discrepancies in the results or attempt to reconcile them, it is reasonable to believe that such events may provide a learning opportunity. Furthermore, the needs of this student in this circumstance are not based on a model of how the student interacts with the pedagogical content of the assignment.

We suggest that consistency across several dimensions may be dynamically monitored and used to adaptively control scaffolding (additional guidance) for the student. Scaffolding involves (a) reducing the number of steps required to solve a problem by simplifying the task (b) keeping the student motivated (c) marking discrepancies between actions taken and the desired solution (d) controlling frustration and (e) demonstrating an idealized version of the task [9]. It is a technique used primarily when students are learning new material, and cannot handle complex problems. Graesser et al showed that use of scaffolding, including good questions and answers, could promote deep inquiry, which students tended to avoid without prompting [10].

For example, suppose students did not follow the directions on the worksheet correctly and did not realize it. An ideal student would have recognized that “something was wrong” and taken some action to resolve the cognitive dissonance, checking the last command executed or re-entering the command. If still confused, the student could ask a classmate or the instructor for help. As part of an online assessment system, we wish to encourage such behavior. An appropriate first step would be to emphasize the cognitive dissonance. In the assignment in the experiment, expectations are outlined in the text, e.g., by writing “Apply a formula that makes a monochrome image in which the cedar foliage is white and everything else is black” before providing the formula. However, the discrepancy could be highlighted further by showing an image of the expected result and asking the student, “Does your image look like this?” before proceeding.

In the extreme case when the same error is repeated, we can assume that the mistake is not inconsistency but represents some higher order misconception. For example, in the PixelMath interface there are buttons for common functions, such as XOR. Other functions can be typed in the window. One command asked students to apply the BXOR formula to exclusive-OR two images on a bit-by-bit basis, and many incorrectly applied the XOR function. It is possible that students who do not correct their error with appropriate scaffolding may not understand the difference between bitwise-exclusive-OR and straight exclusive-OR, or may not realize that they can type equations directly into the PixelMath formula area without clicking buttons. This might require specific formative feedback that reteaches these concepts to the student.

A common misconception about image processing systems such as PixelMath is that the formula tells where to move each pixel of the source image. In reality, the formula describes, for each destination pixel location, where or how to get its value. This “push-instead-of-pull” notion is exhibited by students asked to come up with a formula to, say, reduce the size of an image by a factor of two. Instead of writing Source1(x*2, y*2)

which is correct, they write Source1(x/2, y/2). Similarly, to shift an image 5 pixels to the left, they should write Source1(x+5, y), but they put down Source1(x-5, y). After seeing a consistent pattern of such incorrect formulas, an automatic feedback system could provide scaffolding to specifically address the "push-instead-of-pull" misconception.

6 Conclusions

We have identified a dimension of student performance, consistency, that is content independent, easily mined from educational logs, and that is related to performance outcomes. We suggest that because consistency is an assumption that underlies many educational interventions, the significance of lack of consistency may be overlooked as a potential opportunity to provide scaffolding. We give some suggestions for how an adaptive learning system might exploit consistency measures to scaffold instruction, and to identify when consistency of incorrect responses suggests moving to an intervention based on more sophisticated models of the learner, content, and their interaction.

References

[1] R. Abelson, *Theories of cognitive consistency: a sourcebook.*, Chicago: Rand McNally, 1968.

[2] V.J. Shute, "Focus on Formative Feedback," *Review of Educational Research*, vol. 78, Mar. 2008, pp. 153-189.

[3] R.L. Bangert-Drowns, C.C. Kulik, J.A. Kulik, and M. Morgan, "The Instructional Effect of Feedback in Test-Like Events," *Review of Educational Research*, vol. 61, Jan. 1991, pp. 213-238.

[4] D. Sleeman, A. E. Kelly, R. Martinak, R. D. Ward, and J. L. Moore, "Studies of diagnosis and remediation with high school algebra students," Jul. 2005.

[5] K. Scalise, T. Madhyastha, J. Minstrell, and M. Wilson, "Improving Assessment Evidence in e-Learning Products: Some Solutions for Reliability," *International Journal of Learning Technology (IJLT)*, In press. .

[6] P. Legree, J. Psotka, T. Tremble, and D. Bourne, "Applying Consensus-Based Measurement to the Assessment of Emerging Domains," Jan. 2005.

[7] J. Psotka, "Psychophoresis and Intelligent Self Assessment (ISA) Scales," *Annual Meeting of the Society for Intelligence Research (ISIR)*, Decatur, GA: 2008.

[8] T.M. Madhyastha, "The Relationship of Coherence of Thought and Conceptual Change to Ability. ," *Annual Meeting of the American Educational Research Association*, San Francisco: 2006.

[9] J. Bransford and National Research Council (U.S.).;National Research Council (U.S.)., *How people learn : brain, mind, experience, and school*, Washington D.C.: National Academy Press .

[10] A.C. Graesser, D.S. McNamara, and K. VanLehn, "Scaffolding Deep Comprehension Strategies Through Point&Query, AutoTutor, and iSTART.," *Educational Psychologist*, vol. 40, Fall2005. 2005, pp. 225-234.

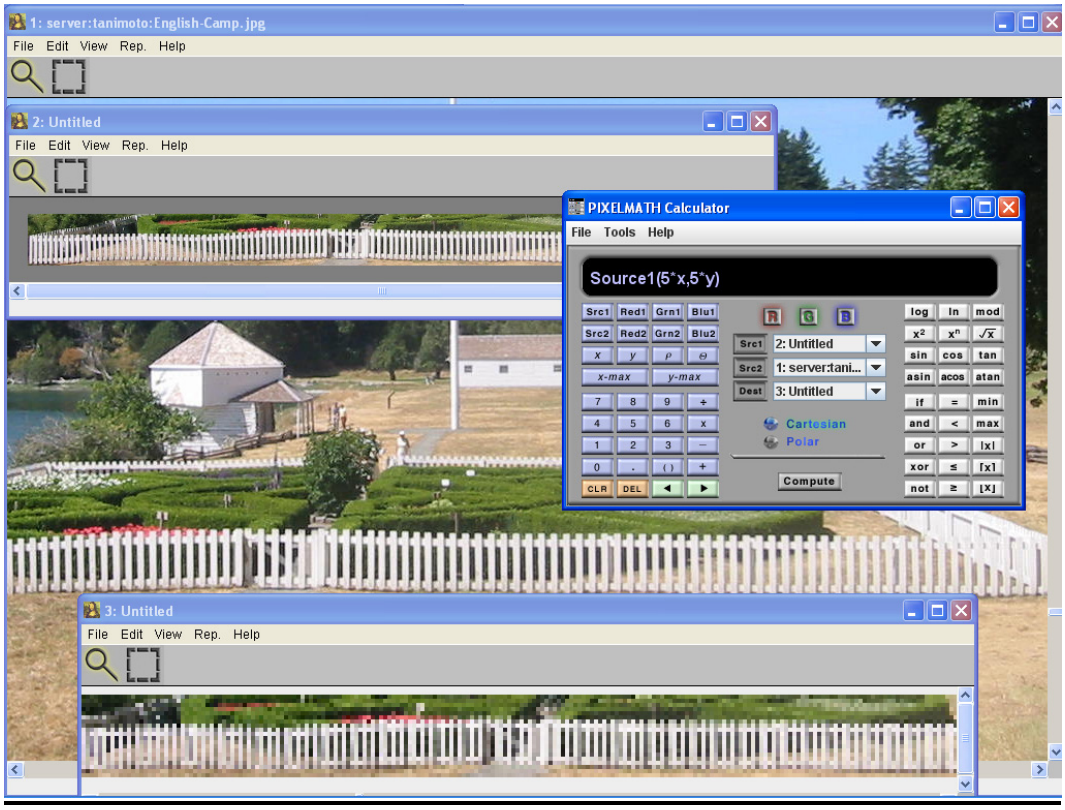

Figure 1. The background window contains the original image used by students for the laboratory activity. The fence has been extracted into another window (above, and zoomed out by a factor of 2) and downsampled into another (below). The PixelMath calculator can also be seen here with the correct formula for the downsampling: Source1(5*x, 5*y).

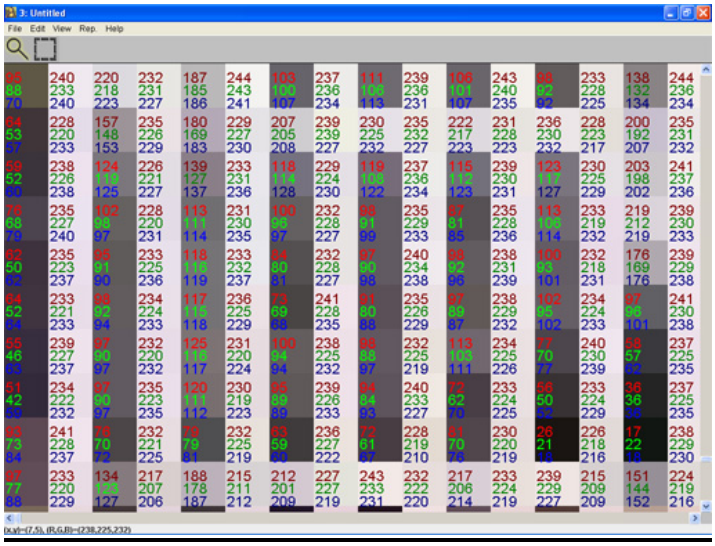

Figure 2. Detail within the downsampled picket fence showing the effect of sampling at the Nyquist rate. Also, PixelMath's display of the RGB values of each pixel can be seen.