Students' Rating of Problem Similarity as a Measure of Problem Solving Expertise

Fran Mateycik[1], N. Sanjay Rebello[1], and David Jonassen[2]

[1]Kansas State University

[2]University of Missouri, Columbia

Abstract

Recognizing the deep structure differences and similarities between problems has been shown to be an essential mark of expertise in problem solving. While novices focus on surface features of a problem, experts have been shown to focus on deep structure. We report on a year-long study with students participating in a treatment to facilitate expert like problem solving. To assess development of student problem solving expertise, students toward the beginning and the end of the treatment were asked to rate the similarities between problem pairs. We report on the results from the similarity ratings of these students. We will present a comparison of the students before and after the treatment as well as compare the student similarity ratings with those of physics faculty members.

Introduction

Problem solving is regarded as an important cognitive skill that all people must develop (Hsu, Brewe, Foster, & Harper, 2004; Jonassen, 2000). Our overarching study focuses on case reuse, a process of solving problems by using what was learned through similar previously solved problems (Faltings, 1997). Case reuse is based on the premise that students construct or modify a previously defined conceptual schema by analyzing a worked example. This mental organization of knowledge is then retrieved while solving similar problems. For a schema to be useful in problem solving it must be tied to the inherent physical principles, or deep structure, of the problem rather than its surface features.

Our overall goal was to facilitate the development of conceptual schema by enabling students to focus on deep structure of problems. During this project, we looked at assessing whether our treatment affected students' identification of deep-structure similarities between problems using a similarity rating task. Students participating in our study were given the task of rating the similarity between pairs of problems of varying similarities in surface features and deep structure features.

We address the following research questions in this paper:

Q1) To what extent do students attend to the surface or deep-structure features in comparing problems?

Q2) How do students' ratings of similarity between problems compare between separate treatment groups?

Q3) How do faculty ratings align with theoretical expected ratings?

Q4) How do students' ratings of similarity between problems compare with faculty members' ratings of the same problems?

Research suggests that learners fail to recall examples or schema appropriately because their retrieval is based upon surface similarity between cases, not their deep structural features (Catrambone & Holyoak, 1989; Chi, Feltovich, & Glaser, 1981; Reed & Bolstad, 1991). Catrambone and Holyoak also suggest that generalization improves when problems emphasize structural features shared with a similar example. Research by Chi, Siler, Jeong, Yamaguchi, & Hausmann (2001) has shown that students tend to group problems based on surface features, while experts group problems based on their deep structure. Our tasks were different from those presented by Chi in her research. Rather than ask students to categorize the problems we presented students with pairs of problems and asked them to rate the similarity of each pair on a five-point Likert scale with ’0’ labeled as ‘completely different’ and ‘5’ labeled as ‘identical.’

Methodology

Ten students participated in eight, 75-minute long, focus group learning interview sessions during the Spring of 2008. These students were referred to as the Phase II participants, as they were a part of a larger project’s second phase. Twelve students participating in six, 75-minute long focus group learning interview sessions during the Fall of 2008 were called Phase III participants. Both student groups were representative of the class demographic profile. The topic in each session followed those currently being covered in the algebra-based physics class all participants were enrolled in.

During each focus group learning interview session, a moderator handed out a fully solved example problem and a pair of problems for students to work and analyze. The example problem provided was comparable in physical concept and principle (e.g. energy and ‘conservation of energy’, respectively) to the unsolved problems. All three problems also contained different surface features, (e.g. auto-mechanic’s hydraulic lift, a barrel, and a swimming pool). Participants were asked to individually solve one of the two unsolved pairs using the solved example for assistance. Students were then asked to describe their solution with another participant who was assigned the other unsolved problem. After students discussed their solutions with their partner, they were asked to discuss the similarities and differences between each of the problems, including the solved example.

To assess the impact of using direct deep-structure similar problem comparison during the group learning interviews, the students from both Phase II and Phase III were also required to participate in individual interview sessions. The Phase II participants were required to participate in two individual interviews, one toward the middle and the other toward the end of the semester. The Phase III participants were only required to participate in an end-of-semester individual interview due to time constraints brought on by the spring course schedule. There were several tasks used to assess changes in problem solving proficiency, though one task in particular asked students to rate the similarities between several pairs of problems. The problem pairs were constructed from problems that had facial (i.e. surface) similarities and differences as well as principle (i.e. deep structure) similarities and differences. The overarching concept remained the same across all problem sets, thereby leaving some basic similarity among all problems. This task measured the perceived significance of ‘limited’ variance of problem structure. Researchers created problem pairs with four basic combinations of facial/principle similarities/differences. These are labeled problem pair types A, B, C and D in Table 1. It would have been preferable to create more possible degrees of variance in surface and deep-structure similarity, but the task would have required substantially more time than that available with our student cohorts.

Each student was presented with eight pairs of problems. Students were presented the problem pairs in order A, A, B, B, C, C and D, D. Students were not allowed to backtrack and change their similarity rating for any pair until the end of the sequence when they were given the opportunity to review their ratings for all pairs and decide whether they wanted to revise any of the similarity ratings. Figure 1 shows examples of the similarity rating tasks used in the study in Phase II, Interview 1.

Four non-PER (Physics Education Research) faculty members were also asked to complete the similarity ratings task at the end of the Spring 2008 semester. These faculty members were either currently teaching or had recently taught an introductory physics course. PER faculty were not asked to complete this rating due to their familiarity with this project. Data were collected from these faculty members to compare to our hypothesized expert ratings. This data were also compared to student data on similarity ratings, though statistical significance could not be shown due to the small sample size. We expected the faculty members to be most sensitive to the principle similarities and differences, rather than facial similarities and differences. Thus, we expected that the faculty members would rate problem pairs A and B as ‘high’ on the Likert scale since they both shared principle similarities while rating pairs C and D ‘low’ on the Likert scale, because they both had principle differences. More specifically, we would expect problem pair type D to rate lowest overall and problem pair type A to rate highest overall.

Results

We averaged the similarity ratings of each student in Phase II and III for each problem pair type for each interview. Results for Phases II and III are presented in this paper below.

We also averaged the similarity ratings of each faculty for each problem pair type. The faculty ratings were then used to determine whether the actual faculty ratings aligned with how the researchers expected the faculty to rate the problem pairs. Previous research (Chi, et al., 1981; Chi, Siler, Jeong, Yamaguchi, & Hausmann, 2001) suggests that expert physics problem solvers emphasize physical principles over facial features. It would be expected then that problem pair type A would rate highest with both principle and facial similarity, problem pair type B would rate second highest with principle similarity and facial differences, problem pair type C would rate third highest (or second lowest) with facial similarity and principle differences and finally problem pair type D would rate lowest with principle and facial differences. After faculty ratings were averaged, it was apparent that the faculty did in fact rate the problem pairs as expected.

There was not enough faculty data to warrant any statistical calculations, but Phase II and Phase III end interview averages were also compared to faculty averages. Those informal comparisons are also presented in this paper below.

Phase II Results

Interview 1 was conducted after students completed the first four focus group learning interview sessions. However, the protocols for these interviews were not finalized until the fourth interview, so students were not participating in activities that required them to explicitly focus and reflect on problem similarities and differences. Figure 2 shows principle/facial similarities/differences in each type (P=Principle, F=Facial, S=Similarity, D=Difference). The error bars are the standard deviation over all students and all problem pairs of a given type.

In our results for interview 1 we find statistically significant differences between the similarity ratings of pairs A and B (p-value 0.000), B and C (p-value 0.003) and C and D (p-value 0.008) using two-tailed t-test analysis. The fact that students have rated pairs B and D as significantly lower than pairs A and C is consistent with the notion that students appear to be focusing on facial similarities and differences rather than similarities and differences in principle. For instance, they rate pair B significantly lower than pair A even though the problems in pair B are only facially different. Similarly, they rate pair C significantly higher than pair D even though the problems in pair C have differences in underlying principle.

Through discussion of the similarity ratings with students during this task, it becomes apparent that students recognize problems are related by conservation of energy, but they believe the differences in facial features have a direct effect on the types of energies involved and these are enough to make the solution that much more different. One student stated, “I guess that both the stone and the piano have potential energy like when they’re starting, but that doesn't matter really. It's a totally different technique used to solve each problem. There’s a spring energy now.” It is also apparent through the conversation that pair C problems are different in terms of the method necessary to solve the problems, but are not ‘significantly’ different. “Except this one you're gonna be using a tiny different equation in the path [solving procedure] than this one and that [part of the solution] was the same.”

Interview 2 was conducted after students completed all eight of the focus group interview sessions. At this point, students participated in five finalized focus group learning interviews. Here we found that the differences between A and B and pair B and C are no longer statistically significant. The only statistically significant difference is between C and D (p-value 0.014). The fact that students are rating pairs A and B at about the same level of similarity is consistent with the notion that students have now begun to recognize that the problems in pair B have principle similarities that overpower their facial differences to the extent that they rate pair B almost the same way as they rate pair A. In other words, it appears from these data that students are emphasizing the similarities in principle although there may be facial differences between the problems in pair B. The ratings for pairs C and D in interview 2 are close to identical to their ratings for these pairs in interview 1. We would be interested in seeing the rating for pair C to be significantly less than before and as low as the rating for pair D. Such data would have been consistent with the notion that students are able to overlook the facial similarities in pair C and recognize the difference in principle, but our data do not appear to show this pattern. Rather, it appears from our data that when shown a problem pair that is facially similar, students do not probe further to reflect on whether or not these problems' similarities/differences in principle are significantly impacting the solution.

Data collected from four faculty members at the same institution were also compared to the data collected from students in phase II, interview 2. There were not enough faculty to warrant any statistical calculations, but we can see from the small sample that those faculty participants agree with the ideal hypothetical expert. Problem pair types A and B are both rated high and close to one another, while problem pair types C and D rate lower and close to one another. Figure 3 below shows the average rating for each problem pair type given by faculty. It can also be seen in Figure 3 that students' ratings for three of the four problems are similar to the faculty' ratings by the end of the semester. Problem type C is most different. Students rate type C problem pairs higher than type A and B problem pairs, while faculty rate type C problem pairs lower than type A and B problem pairs.

Phase III Results

For the similarity rating task used during the Phase III interview, students rated the same four types of pairs at the end of the semester following the focus group learning interview treatment. Students rated pair types A, B and C all at about the same level of similarity, remaining consistent with the previous semester. Unfortunately, these results cannot be shown significant through statistics as the student populations are different. In hindsight, it would have been much more useful to have a baseline or control group to compare data between students of equal variance, but we were unable to collect sufficient volunteers for a baseline in the spring or fall semester. Figure 4 shows the mean ratings for each problem type for the individual interview conducted in Phase III and the individual interview 2 conducted the previous semester.

Students were also asked to discuss their ratings with the interviewer. Similar to Phase II, students noted that the problems given in all pairs shared the same concept.

> “Well, all of these problems are alike. They are all problems involving simple harmonic motion and I could rate them all very close. In fact I did!.....This pair and this pair (last two pairs, type D) are still similar to the rest, but they require just a little more work, so they got a slightly lower number than the rest.”

Four faculty were also asked to rate the problem pairs. Their average ratings were plotted over the top of the end interview ratings from Phase II and Phase III as seen in Figure 5. Phase II and III participants rated pair types A, B and D all at about the same level of similarity as compared with the faculty ratings. The most significant difference between the students’ final interview ratings and the faculty ratings is the rating of problem pair type C. Problem pair type C is rated lower than problem pair types A and B for faculty. Students rate problem pair type C higher than problem pair types A and B. This difference in ratings suggests that students deemphasize facial (surface) features ONLY when problem pairs are not facially similar. In problems which share facially similarity, these elements dominate student similarity ratings.

Conclusions

We address each of our research questions below:

Q1) To what extent do students attend to the surface and deep-structure features in comparing problems?

This question may only be answered with respect to the Phase II cohorts because they were asked to complete two individual interviews. Before our focus group learning interviews, students in the Phase II cohort rated problems sharing prominent surface features higher than problems with different surface features. After our focus group learning interviews, students' ratings of problems sharing surface features remained high, but problems with different surface features and similar deep-structure features were also rated high. A statistical significance could not be determined because problems used for interview 1 were not the same as those used in interview 2 and each set concentrated on different topics covered during the semester. Students rated pairs A and B at about the same level of similarity. These data were consistent with the notion that students recognize problems in pair B as having principle similarities that overshadow their facial differences to the extent that they rate pair B almost the same way as they rate pair A.

Q2) How do students' ratings of similarity between problems compare between separate treatment groups?

Phase II and III participants rated pair types A, B and C all at about the same level of similarity. Unfortunately, these results cannot be shown significant through statistics as the student populations are different.

Q3) How do faculty ratings align with theoretical expected ratings?

The four actual faculty ratings aligned with how the researchers expected the faculty to rate the problem pairs. Problem types A, B, C and D were rated highest to lowest in consecutive order.

Q4) How do students' ratings of similarity between problems compare with faculty members' ratings of the same problems?

A direct comparison is difficult with such small numbers, but if we look at general trends, the faculty' ratings and students' ratings after treatment are very close for problem pair types A, B and D. Problem pair C, which includes problems that are facially similar and principle different, are rated lower than problem pair types A and B for faculty, but higher than problem pair types A and B for students. Students learn to deemphasize facial features when given problems that are not facially similar. When problems share facial similarity, the students no longer attend to the differences in principle between problems.

Implications and Future Work

The change in students' ability to discern the similarities and differences in Phase II interviews 1 and 2 could be due to not only the participation in the focus group learning interviews. They could also be due to the differences in the specific problems used in each interview and/or the topic on which they were based. Students for both Phase II and Phase III were simultaneously enrolled in an algebra-based physics course which also could have altered the deep-structure feature emphasis on these problem similarity rating tasks. Nevertheless, this work provides promising evidence that explicit contrasting of examples and problems can significantly impact the perceived importance of deep-structure elements in a problem statement.

Future work is currently being conducted with a larger population of algebra-based students. The new study integrates the explicit contrasting of similarities and differences into the curriculum of the course homework and laboratory activities. The similarity variance between problem pairs could be further augmented to include greater differences in structure and problem representation.

Acknowledgements

This research is supported in part by NSF grant DUE-06185459.

References

Catrambone, R., & Holyoak, K. (1989). Overcoming Contextual Limitations on Problem-Solving Transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition , 15* (6), 1147-1156.

Chi, M., Feltovich, P., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science , 5* (2), 121-152.

Chi, M., Siler, S., Jeong, H., Yamaguchi, T., & Hausmann, R. (2001). Learning from Human Tutoring. *Cognitive Science , 25*, 471-533.

Faltings, B. (1997). Case Reuse by Model-Based Interpretation. In M. Maher, & P. Pu (Eds.), *Issues and Applications of Case-based Reasoning in Design.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hsu, L., Brewe, E., Foster, T., & Harper, K. (2004). Resource Letter RPS-1: Research in Problem Solving. *American Journal of Physics , 72* (9), 1147-1156.

Jonassen, D. (2000). Toward a Design Theory of Problem Solving. *Educational Technology and Research and Development , 48* (4), 63-85.

Reed, S., & Bolstad, C. (1991). Use of Examples and Procedures in Problem Solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition , 17*, 753-766.

Table 1

Problem Pairs for the Similarity Rating Task

	Facial Similarity (**FS**)	Facial Difference (**FD**)
Principle Similarity (**PS**)	**A**	**B**
Principle Difference (**PD**)	**C**	**D**

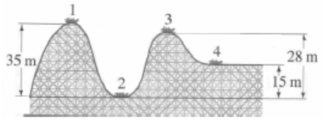
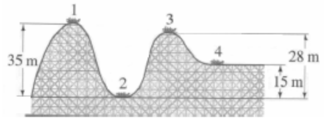
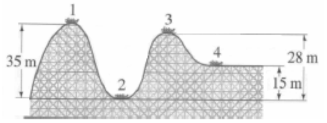
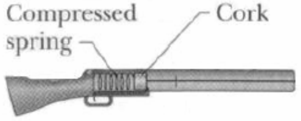
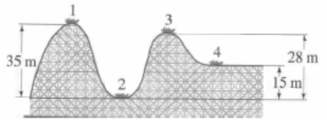
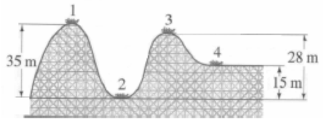
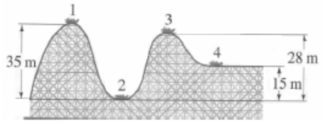
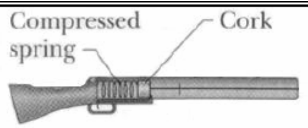
Type A	 An 800 kg roller coaster shown in the figure above is dragged up to point 1 where it is released from rest. Assuming the track is frictionless; calculate the speed at point 3.	 A roller coaster shown in the figure above will be moving with a velocity of 22 m/s at the exact moment it hits point 2. Assuming the track is frictionless; calculate the speed at point 4.
Type B	 An 800 kg roller coaster shown in the figure above is dragged up to point 1 where it is released from rest. Assuming the track is frictionless; calculate the speed at point 3.	 A 0.10 kg bullet is loaded into a gun tilted upward at a 30° angle from the horizontal, compressing a spring (spring constant is 6400 N/m) a distance of 0.20 m. When the trigger is pulled, the spring is released, and the bullet leaves the spring at the spring's relaxed length at a speed of 50.5 m/s. The bullet travels a distance of 0.60 m before exiting the barrel of the gun. What is the speed of the bullet as it leaves the gun?
Type C	 An 800 kg roller coaster shown in the figure above is dragged up to point 1 where it is released from rest. Assuming the track is frictionless; calculate the speed at point 3.	 An 800 kg roller coaster shown in the figure above is dragged up to point 1 where it is released from rest. The work done by friction in going from point 1 to point 3 is 4800 J. Calculate the speed at point 3.
Type D	 An 800 kg roller coaster shown in the figure above is dragged up to point 1 where it is released from rest. Assuming the track is frictionless; calculate the speed at point 3.	 A 0.10 kg bullet is loaded into a gun compressing a spring (spring constant is 6400 N/m) a distance of 0.20 m. When the trigger is pulled, the spring is released, and the bullet leaves the spring at the spring's relaxed length. The bullet travels a distance of 0.60 m before exiting the barrel of the gun. The coefficient of kinetic friction between the bullet and the barrel is 0.10. What is the speed of the bullet as it leaves the gun?

Figure 1. Example problem pair types used during Phase II, Interview 1

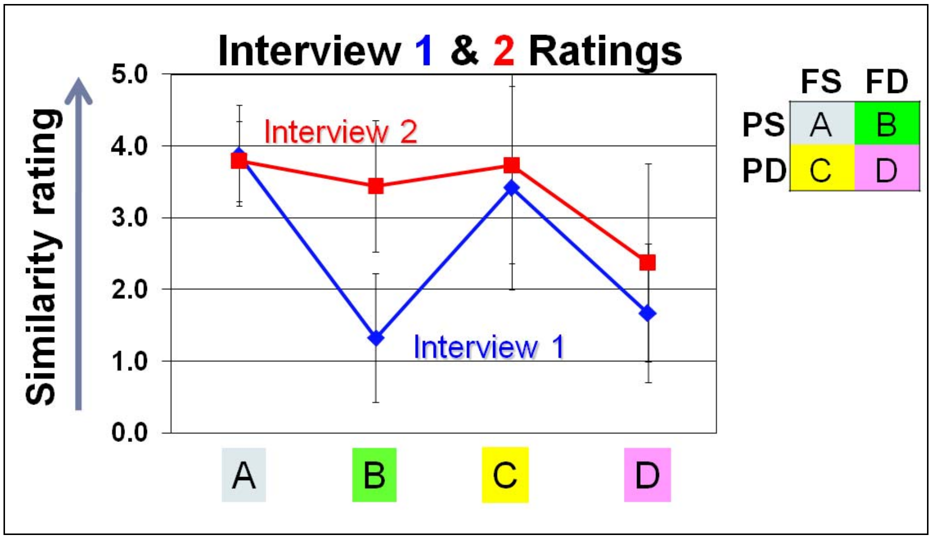


Figure 2. Students' similarity ratings of problem pairs of types A, B, C, and D for Phase II interview 1 and 2

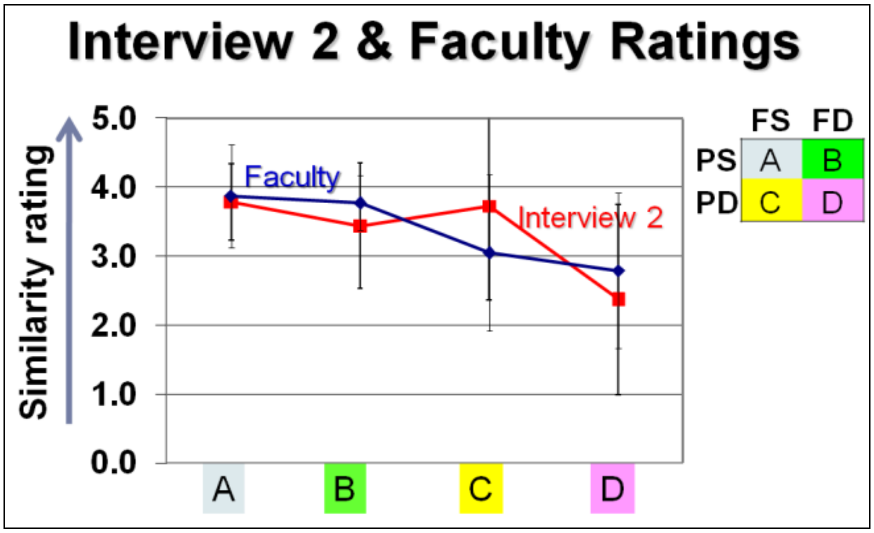


Figure 3. Student ratings from Phase II, interview 2 and faculty ratings

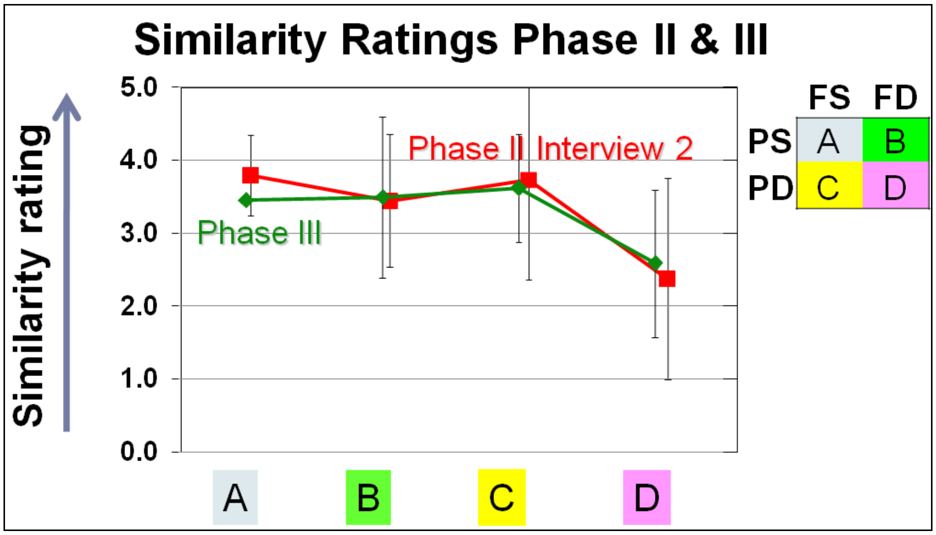


Figure 4. Similarity ratings at the end of treatment for Phase II and Phase III

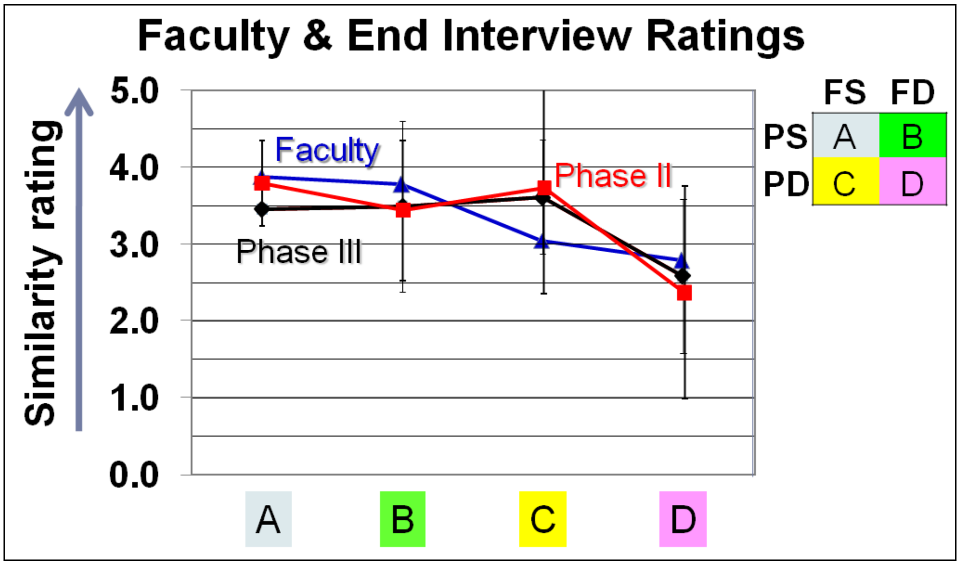


Figure 5. Faculty and end interview ratings from Phase II and Phase III