Judging Similarity Facilitates Deriving a New Solution Procedure

Robert S. Ryan, Ph.D.
Kutztown University

An experiment examined the effect of comparing the solution procedures of algebra word problems during training on understanding and using a general principle at test. Participants either made a similarity judgment regarding the training examples' goals (N = 46) or only solved the examples (N = 98). On transfer problems, the similarity judgers were better able to generate a complex and previously unencountered solution procedure and use it to solve a transfer problem more difficult than the training problems. These results are interpreted as showing that the similarity judgers had developed a better understanding of a general equation that they were not taught.

Problem solving expertise is characterized by knowledge of general principles (Chi, Feltovich, & Glaser, 1981). However, people prefer to learn from examples rather than from instructions about an abstract principle (LeFevre & Dixon, 1986). Fortunately, we know that when people compare multiple examples they are better able to infer a general principle that guides a solution procedure. They can later use that solution procedure to solve a superficially different new problem (e.g., Catrambone & Holyoak, 1989; Cummins, 1992; Gick & Holyoak, 1983). For practical applications, however, it would be helpful to know whether such a comparison process is also useful for generating a *new* solution procedure. Therefore, this paper will examine a hypothesis suggested by previous theories of analogical reasoning. Specifically, comparing the solution procedures of algebra word problems should improve performance not only on superficially different problems that can be solved by the *old* solution procedure, but also on problems that require using a general principle to generate a *new* solution procedure.

The structure mapping approach to analogy suggests that judging similarity enables people to recognize similar relations even when similar surface features are potentially misleading (Gentner, 1983; Markman & Gentner, 1990; 1993). Accordingly, Gick and Holyoak (1983) found a benefit for giving participants training in two example problems that were analogies to a later test problem, rather than just one. The training made explicit that the examples shared a common abstract principle from which their solution procedure was derived. The abstract principle was using a strong force to destroy a target, but applying it in a way that prevented collateral damage. The solution procedure was always to divide the strong force into multiple weaker forces and converge them on the target. Participants trained with two examples were better able to solve the test problem, even though it had new surface features, because they had inferred the relations in the abstract principle behind the solution procedure.

In Gick and Holyoak's (1983) studies the participants never had to use the principle to generate a *new* solution procedure. Sometimes, however, different problems belonging to the same general class can have *different* solution procedures, although those solution procedures can be derived from the *same* general principle. If training in such problems made their *solution procedures* explicit, but left the general principle only *implied*, then would comparing such examples be beneficial? Specifically, would it enable people to generate a *new* solution procedure, and use it to solve a transfer problem that not only had new surface features, but also required the *new* procedure?

Algebra Word Problems With a Common General Principle But Different Solution Procedures

Appendix A shows examples of training problems and test problems, respectively, that were used to address the question above (other Appendixes with more detailed information are posted on the internet[1]). All of the problems were based on the general principle of weighted averaging of ratios. However, the surface features and solution procedures of the test problems were independently either the same as (old), or different from (new), the training problems (see Appendixes B and C).

All of the problems consisted of initial amounts with associated ratios that were combined into a final amount with an associated ratio. The training problems were presented in pairs. For the training pairs, the surface features of the members of the pairs were always different from one another. The solution procedures, which depended on the problem's goal, were sometimes the same for both members of the pair, and sometimes different. These training pairs were used in an experiment that manipulated whether participants only solved the examples or compared them. For the comparison task the participants made a same/different judgment regarding the goals (i.e., the unknown).

The unknown could either be the final ratio (i.e., the "combined" ratio), or one of the two initial ratios. Performing the similarity judgment task required recognizing which of the ratios was the goal for each member of the training pair. In an early part of the training (before participants practiced their task on their own) some of the training pairs were used to explicitly show the participants which ratios were initial and which one was the final. However, the ratios were identified only in terms of a step by step procedure, not in terms of a general equation.

In the step by step procedure, the unknown, of course, always occurs in the last step, in which its value is determined, as shown in Appendix B. Therefore, givens and unknowns always occur in the same part of the procedure, whereas initial and final elements occur in different parts of the procedure for problems with different unknowns. It is the structure of the equation that determines which elements are the initial and final elements, and they always occur in the same location in the equation, regardless of which element is the unknown. Therefore, focusing on which problem element is the initial versus the final element, regardless of whether it is the unknown, could lead to noticing the problem structure in terms of the general principle of weighted averaging that is embodied in the equation.

For a control condition, participants solved the examples as their training task. In the "solve" training, participants were shown a step by step solution procedure. In the "judge similarity" training, they were shown the same step by step procedure, but they were also shown how to distinguish between the initial and final elements. This was done in order to enable them to determine whether the members of the training pairs were the same or different in terms of which problem element was the goal. The initial ratios were described as two ratios that were averaged together by weighted averaging and the final ratios were described as the resulting weighted average. However, no participants (in either type of training) were shown a general equation for weighted averaging, nor was the general principle of weighted averaging explained in any way.

Recognizing Implied Relations.

The comparison task *implied* the general principle of weighted averaging because it suggested how the problem elements could be lined up as if they were in the general equation. However, it never made explicit the most important information conveyed by an equation, which is the relations between the elements. According to Gentner's (1983) structure mapping theory, because the critical higher order relations between the lined up elements would be the same, they could be inferred by this comparison process. One of the most important inferred relations would be the overall equality relation in the implied general equation. Therefore, the structure mapping approach suggests that the comparison task could result in inferring the most important information conveyed by the equation. A test of whether that occurred was provided by the new solution procedure problems. Success at this problem would serve as especially good evidence that the equation had been inferred. It requires solving for one of the amounts in the problem, rather than one of the ratios, a particularly difficult unknown to solve for. It is especially difficult because it requires using an equation in which the unknown appears in two places, and those two places are on opposite sides of the equal sign (Koedinger, Alibali, & Nathan, 1999). Therefore, this problem could not be solved by the trained solution procedure, but could be solved by a variation of it that could be derived from the general equation.

The Hypothesis

Comparing the training examples by making a similarity judgment should enable people to recognize the most important higher order relation in the general principle on which the examples are based, even when the principle is only implied. Recognizing that relation should enable them to derive the general principle and to use it to generate a *new* solution procedure. This should enable people who practice comparing the examples to perform better than people who only practice solving them on a transfer test problem that requires the variation.

Method

Participants

The participants were 157 psychology undergraduate students at the University of Pittsburgh who participated as part of the requirements of their Introductory Psychology course. The data from 13 participants were not included because of procedural errors, leaving 144 participants' data in the analysis. The participants were run in sessions that included from two to six participants.

Materials

The materials consisted of problem pairs for training and problems for pretest and posttest. Sixteen problems pairs for training were constructed so that the members of the pairs could be either different from one another, or the same, in terms of their goals. However, it was necessary to make the surface features of the members always different from one another. When people are asked to make a similarity judgment, they have a strong tendency to base the judgment on surface features (Ryan, 2003). However, the similarity judgment was to be a same/different judgment regarding the training problems' goals. Therefore, it was necessary to make very clear that surface features were not to be used as the basis for the judgment. Furthermore, the purpose of the judgment was to facilitate inferring the one characteristic that was always to be the same, i.e., the underlying principle. Therefore, the best way to make clear that surface features were not to be used as the basis for the judgment, and at the same time to facilitate noticing that *only* the underlying principle was always the same, was to make the surface features always different.

The test problems had either one of the old (trained) procedures or a new procedure. This provided the test of primary interest, that is, a test of the very difficult task of transfer to new problems that would require using a general equation to generate a new procedure. In order to show that such transfer is much more difficult than transfer to new problems with just a different set of surface features, test problems with both the same surface features as that used in training, and different surface features were used. Crossing these factors resulted in four kinds of test problems, as shown in Appendix C. Two tests of equivalent difficulty, each test containing one each of the four kinds of test problems, were constructed. One was used as a pretest and one as a posttest, with the order counterbalanced across participants.

Procedure

The procedure consisted of a pretest, a three part training session, and then a posttest. In the pretest the participants were presented with one each of the four kinds of test problems in a random order. They were allowed three minutes to work on each problem.

The three-part training procedure consisted of worked examples, guided practice, and unguided practice. The tasks during training differed depending upon whether the participants were in the experimental condition, called the judge-similarity condition, or in one of two control groups. The task for the participants in the judge-similarity condition, was to write a short explanation of whether a training pair was a same or a different pair (Appendix B provides examples of their explanations). The task for the control participants was to solve the same training problems, but not to judge their similarity.

Worked examples. The two pairs of training problems shown in Appendix B were used as worked examples. All participants received the same two training pairs in the same order. The experimenter allowed the participants to read through the worked examples one at a time. The experimenter also gave a brief oral explanation of each one after the participants read them and answered any questions to make sure the participants understood the examples. This procedure usually took 10 to 15 minutes.

For the judge-similarity participants (N = 46), and for half of the control participants (called the solve-with-instructions participants[2]) (N = 49), the worked examples were not just samples of how to solve the problems, but also samples of how to judge similarity. Appendix D shows the training materials for how to judge similarity that were added to the worked example problems for those two conditions. The other half of the control participants (called the solve-only participants) (N = 49) only received worked examples showing how to solve the problems.

Guided practice. Two more training pairs were used for guided practice. All participants received the same two training pairs in the same order. The first pair was a same pair in which finding the initial ratio was the goal. The second pair was a different pair in which the goal of the first member was to find the initial ratio and the goal of the second member was to find the final ratio. The guided practice was untimed. The experimenter guided the participants until it appeared they had all reached the correct solution. This usually took about 10 minutes. In those few cases in which a participant failed to solve the problem correctly, even with the guidance, the experimenter simply told the participants the correct answer. This was done for both the solution to the problems and, in the appropriate conditions, for the judge similarity task.

Unguided practice. After the guided practice, the participants spent 15 minutes in unguided practice on the remaining 12 pairs of problems. The pairs were a random mixture of same and different pairs with the constraint that there were different types of pairs within the first three pairs. This was done because it was expected, based on pilot studies, that some participants, particularly judge-similarity participants, would only be able to do as few as three pairs in the allotted time, and it was necessary for them to encounter both types of pairs. In fact, it turned out that participants in the judge-similarity condition actually did between 3 and 8 problem pairs, with a median of 4.

The judge-similarity participants used these pairs of problems to practice doing the judge similarity task. Both the solve-with-instructions and the solve-only participants used them only to practice solving the problems. Thus, at this point in the training, the participants either solved the problems or engaged in the judge similarity task, depending on their condition, but no participants did both. Solving a pair of problems took less time than judging the similarity or difference between them. Therefore, the solve-with-instructions and the solve-only participants usually did more problems than the judge-similarity participants in this part of the training. Nevertheless, it provided a controlled amount of time on task for all participants. The procedure concluded with the posttest. The same procedure was used on the posttest as was used on the pretest.

In summary, the participants experienced the following. First, they took a pretest on four different ratio problems. Next, they were trained with pairs of examples. First, they saw two sample pairs which provided worked examples. Next, they received guided practice using two more pairs of problems. Finally, they spent 15 minutes in unguided practice on more pairs of problems. Each of these three parts of the training either trained them in just how to solve the problems or in both how to solve the problems and how to do the judge similarity task depending upon their condition. After the pretest and training, all participants took the posttest. The entire experiment, including opening remarks, instructions for the various tasks, and debriefing, took just short of two hours.

Design and Analysis

The experiment had a 2 (order of test) by 2 (old solution procedure, new solution procedure) by 2 (old surface features, new surface features) by 2 (judge similarity, solve) design. The new solution procedure or surface features in the design refers to the difference between the test problems and both members of the training pairs, not the difference between members of the training pairs. The solution procedure and surface features factors were within subjects, whereas the training condition factor and the order of test were between subjects. The main dependent variable was performance on the posttest problems scored as either correct or incorrect. Those scores were analyzed with an analysis of covariance in which the pretest performance, scored as either correct and incorrect, was the covariate.

Partial Credit Scoring

It was possible for participants to either follow the trained procedure or to use an equation to an extent that indicated their degree of understanding, but without reaching a correct answer. Therefore, the participants were also given a partial credit score[3] on the pretest and posttest for their use of both the trained solution procedure and the general equation. The procedure and equation scores consisted of different numbers of possible points, but were converted to proportions for purposes of comparison.

Results and Discussion

Performance on the Posttest as a Function of the Different Test Problems.

The overall analysis of performance on the posttest problems showed, as expected, that the problems differed greatly in difficulty. Therefore, the effects of the different kinds of posttest problems collapsed over the training conditions are presented first. It also showed that training condition interacted with solution procedure, but not with surface features. Therefore, the effect of training condition on the old and new solution procedure problems collapsed over surface features is presented next. Finally, the effect of training condition on the use of the trained procedure and the general equation is presented. A significance level of .05 is used for all tests.

	Surface Features		
Solution Procedure	**Old**	**New**	**Total**
Old	66	43	55
New	12	14	13
Total	39	29	34

Table 1. Mean Percent Correct on the Posttest (Adjusted for Pretest Performance) for Each of the Four Kinds of Transfer Test Problems.

As shown in Table 1, performance was much better on test problems with the old solution procedure than on those with a new solution procedure, $F(1, 420) = 178.44$, $p < .001$, $MSE = .11$. Performance was also better on old surface feature problems than on new surface features, $F(1, 420) = 12.04$, $p = .001$, $MSE = .11$. Among the old solution procedure problems, the advantage for old surface features held as a simple effect, $F(1, 420) = 30.29$, $p < .001$, $MSE = .11$. However, among the new solution procedure problems, performance was slightly better when the problem had new surface features, although this simple effect was not significant, $F(1, 420) < 1$. The result was a significant interaction between solution procedures and surface features, $F(1, 420) = 19.10$, $p < .001$, $MSE = .11$. However, this interaction should be interpreted cautiously because the apparent lack of effect of surface features when the solution procedure was new may have been a floor effect.

Performance on the Old and New Solution Procedure Problems as a Function of Training Condition

As shown in Table 2, the judge similarity participants performed better than the solve participants on the test problems with a new solution procedure. As a simple effect, this benefit was significant, $F(1, 559) = 4.13$, $p = .043$, $MSE = .11$. On the other hand, for test problems with the old solution procedure, the solve participants performed better than the judge similarity participants. This simple effect was also significant, $F(1, 559) = 7.19$, $p = .008$, $MSE = .11$. The result was a solution procedure by training condition interaction, $F(1, 420) = 11.49$, $p = .001$, $MSE = .11$. Because of the nature of the interaction, there was no main effect of training condition, $F(1,139) < 1$. There were no other interactions involving training condition, solution procedure, or surface features.

	Training Condition		
Solution	**Judge Similarity**	**Solve**	**Total**
Old	49	60	55
New	17	9	13
Total	33	34	34

Table 2. Mean Percent Correct on the Posttest (Adjusted for Pretest Performance) for the Solution Procedure by Training Condition Interaction.

Effect of Training Condition on Use of the Trained Procedure and Equation

The superiority of the similarity judgers on the new procedure problems was expected because judging similarity was hypothesized to lead to greater use of the general equation. However, the superiority of the solvers on the old solution procedure problems was not expected. It might be explained by the solvers' making better use of the trained procedure. If that is the case, however, then this raises a question about the expected result. Were the similarity judgers' superior on the new procedure problems because of greater use of the equation, or should that result be interpreted as the solvers' being inferior on the new procedure problems due to trying to use the trained procedure on problems where it would be detrimental?

The reason for this question is related to the difference between the similarity judgers and the solvers in the amount of experience with using the trained procedure. During the training, all of the participants received training in a procedure to solve the training problems. This procedure would be useful for solving the old solution procedure test problems, but would not be useful, and could even be detrimental to solving the new solution procedure problems. But using the equation *would* be useful for the new solution procedure problems. On the one hand, experience with the general equation was the same for both groups in that no participants received training in the equation. However, experience with the procedure differed for the two groups. Although all of the participants received training in the procedure, the solve participants practiced using it for their 15 minute unguided practice session, whereas the similarity judgers practiced judging similarity. This raises two possible explanations for the superiority of the similarity judgers over the solvers on the new solution procedure problems. The similarity judgers could have performed better due to greater use of the equation. But the solvers could have performed worse due to trying to use the trained procedure to a greater extent than the similarity judgers. This possibility is suggested further by the fact that the solvers were superior to the explainers on the old solution procedure problems. Because the old solution procedure problems benefit from use of the trained solution procedure, and it was the solvers who were significantly better on these problems, the solvers may have benefited by a greater tendency to use the trained procedure on those problems. If they used it to a greater extent on the old solution procedure problems, then they may also have *tried* to use it (to their detriment) on the new solution procedure problems. In order to determine if this was the case, we examined the differences in use of the trained procedure and use of the equation between the training conditions. The partial credit scoring of the use of the trained solution procedure and the use of the general equation on the posttest problems were analyzed in the same way that the success at solving the posttest problems had been analyzed, that is, as a function of training conditions and solution procedure, collapsed over surface features.

	Training Condition		
Solution Procedure	**Judge Similarity**	**Solve**	**Total**
Old	77	79	78
New	29	17	23
Total	53	48	51

Table 3. Mean Percentage of Use of Trained Procedure on the Posttest (Adjusted for Pretest Performance) as a Function of Training Condition, and Solution Procedure.

Use of the trained procedure. If the solvers' superiority on the old solution procedure problems was caused by their use of the trained procedure, then we would expect to see that they had used it to a greater extent than the similarity judgers on those problems. As can be seen in Table 3, the solvers in fact did *not* use the trained procedure to a greater extent than the similarity judgers. First, on the old solution procedure problems there was no difference between the use of the trained procedure between the solvers and the similarity judgers, F (1, 559) < 1. Second, on the new solution procedure problems, the similarity judgers actually used the potentially detrimental trained procedure more than the solvers, F (1, 559) = 10.10, p = .002, MSE = .09. This resulted in a significant interaction between training condition and solution procedure, F (1, 420) = 6.88, p = .009, MSE = .09.

Apparently, factors other than use of the trained procedure must have contributed to the solvers' poor performance on the new solution procedure problems. But this raises the question of why the similarity judgers' use of the trained procedure on the old solution procedure problems was just as great as that of the solvers in spite of their poorer performance on those problems. Due to manner in which the use of the procedure was scored, higher scores mean that participants successfully progressed further into the procedure, even if they did not finish, and, therefore, did not correctly solve the problem. A careful examination of the distribution of those procedure use scores *among those who were incorrect on the old solution procedure problems* showed two reasons for the seemingly contradictory result. First, among the similarity judgers, more than among the solvers, there was a tendency to progress further into the procedure, even when the progress did not reach a perfect score (and therefore did not reach a correct answer). Second, there was also a greater tendency among the similarity judgers to progress all the way to a perfect score on using the procedure, but to nevertheless still fail to provide a correct answer. Both of these factors contributed to a higher mean procedure use score for the similarity judgers in spite of not contributing to a higher mean solution performance score.

The first of these two factors could indicate that among the participants who were incorrect on the old solution procedure problems, at least the similarity judgers were working faster than the solvers, and were therefore less likely to run out of time before solving the problem. The fact that the similarity judgers were more likely to achieve a perfect score for use of the procedure, and yet still fail to solve the problem, could indicate that, along with using the procedure more quickly, they were using it less carefully. This interpretation is further supported by an examination of the reasons for these seemingly difficult to explain failures. The greatest difference between the groups was that the similarity judgers were more likely than the solvers to make a mechanical error (e.g., an arithmetic fact error)[4].

Thus, there were two differences between the similarity judgers and the solvers. The similarity judgers worked the procedure more quickly, but less carefully, than the solvers. This could indicate that having less experience with the solution procedure during training led to a shallower understanding of the procedure among the similarity judgers.

Use of the general equation. The relative use of the equation by the similarity judgers and the solvers on the new solution procedure problems should also shed light on the true cause of the similarity judgers' superiority on those problems. Table 4 shows that, as would be expected because the new solution procedure problems virtually require the use of the general equation if it is to be solved correctly, the use of the equation was greater for the new solution procedure than for the old solution procedure problems, F (1, 420) = 40.15, p < .001, MSE = .08. There was no main effect of training condition nor interaction, F's < 1. Although the differences between the similarity judgers and the solvers for individual problems were small and not statistically significant, the similarity judgers tended to use the equation numerically more than did the solvers. In fact, for one type of problem (i.e., the new solution procedure, old surface features problem) the percent of use of the equation for the similarity judgers was 31, compared to 23 for the solvers. As a one-tailed t-test in the predicted direction, this difference approached significance, t (559) = 1.63, p = .051. Although this result is presented because it provides at least a suggestion that judging similarity may have led to greater use of the equation, it must be interpreted cautiously because there was no three way interaction between surface features, solution procedure, and training condition to justify comparing training conditions within an individual problem.

	Training Condition		
Solution Procedure	**Judge Similarity**	**Solve**	**Total**
Old	11	9	10
New	29	25	27
Total	20	17	28

Table 4. Mean Percentage of Use of Equation on the Posttest (Adjusted for Pretest Performance) as a Function of Training Condition, and Solution Procedure.

These results raise the question of why the similarity judgers outperformed the solvers on the new solution procedure problems in spite of greater inappropriate use of the trained procedure and the weakness of the evidence for their greater use of the general equation. In a previously reported analysis of the data regarding equation use, Ryan and Schooler (2001) found that, among participants who avoided inappropriately using the trained procedure and who also correctly set up the general equation, the similarity judgers were more likely than the solvers to then go on to succeed at correctly solving the problem. The solvers, who were more likely to fail to solve even though they were making a good attempt to use the equation and were not using the inappropriate procedure, failed dramatically more often because of the mechanical errors. This could have been because the solvers had a less automatized understanding of the equation, and therefore needed to spend more cognitive effort on working with it, thus resulting in simple errors that otherwise would not have occurred.

In summary, there is no evidence in the data to support the alternative explanation that the similarity judgers were superior to the solvers on the new solution procedure problems because the solvers had a greater tendency to inappropriately use the trained procedure. Rather, the similarity judgers seem to have developed a deeper understanding of the general principle of weighted averaging. Similarly, the solvers seemed to have been superior on the old solution procedure problems because they used the procedure more carefully, although more slowly.

This interpretation is supported in the following two ways. First, it is supported by the numerically greater use of the equation on the part of the similarity judgers. Second, it is supported by the findings from Ryan and Schooler (2001) showing that the similarity judgers were better able to *successfully use* the equation than the solvers, even though they did not invoke it significantly more often.

Conclusions

The results of this study show that comparing examples, specifically by means of making a similarity judgment regarding the problems' solution procedures, results in inferring a general principle that is useful not only for applying a trained procedure to a problem with new *surface features*, but also for generating a *new solution procedure* from the general principle. Additionally, the partial credit data suggests that, on the one hand, practicing solving led to using the trained procedure more slowly and carefully, and therefore, to superior performance on problems that could be solved by the trained procedure. But, on the other hand, it also suggests that judging similarity led to a deeper understanding of the general principle and better use of the general equation, and therefore, to superior performance on problems that required generating a *new* solution procedure.

References

Catrambone, R. & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1147-1156.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Cummins, D. D. (1992). Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1103 - 1124.

Gentner, D., (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1-38.

Koedinger, K. R, Alibali, M. W., & Nathan, M. J. (1999, May). *A developmental model of algebra problem solving: Trade-offs between grounded and abstract representations.* Paper presented at the meeting of the American Educational Research Association (AERA) Conference, Montreal.

LeFevre, J. & Dixon P. (1986). Do written instructions need examples? *Cognition and Instruction, 3,* 1-30.

Markman, A. B., & Gentner, D. (1990). Analogical mapping during similarity judgments. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 38-44), Cambridge, MA.

Markman, A. B., & Gentner, D. (1993). Splitting the difference: A structural align ment view of similarity. *Journal of Memory and Language, 32,* 517-535.

Ryan, R. S. (2003, May). *Misleading Surface Features Impair the Recognition of a Useful Analogy.* Paper presented at the annual meeting of the American Psychological Society, Atlanta, GA.

Ryan, R. S., & Schooler, J. W. (2001, June). *Analogical Reasoning With Examples Facilitates Inferring General Principles in Spite of the Risk of Negative Transfer.* Paper presented at the annual meeting of the American Psychological Society, Toronto, ON.

Author Note

Robert S. Ryan, University of Pittsburgh.

The author wishes to thank Jonathan Schooler for comments on earlier versions of this manuscript. Thanks also to Jerry Casale, Barb Pentz, and Cheryl Smith for helping to collect and score the data. Thanks to Morgan Becker for help in preparing Appendix B.

Correspondence concerning this article should be addressed to Robert S. Ryan who is now at the Department of Psychology, Box 730, Kutztown University, Kutztown, Pennsylvania, 19530. Telephone (610) 683-4457, Fax (610) 683-4467. E-mail: rryan@kutztown.edu.

Footnotes

1 To save space, Appendix A provides just two examples each of the training and test problems. Some readers may be satisfied to extrapolate from those examples to understand the more complex details to be described later. However, for more detailed information about the training and test problems, as well as other details about the experiment, Appendixes B, C, and D, have been posted on the internet. Appendix B provides examples of pairs of training problems, their equations, the alignment of their givens and unknowns, the alignment of their initial and final elements, and examples of participants' similarity judgments. Appendix C provides examples of all of the types of test problems and their equations. Appendix D provides the instructions (including an example) for the explain similarity task for a same and a different pair. The URL for those appendixes is: http://faculty.kutztown.edu/rryan/judging_similarity/.

2 The original intention had been to train all the participants to solve the problems, and then to manipulate problem comparison by having the experimental participants spend only the 15 minutes of unguided practice comparing training problems and having all the control participants spend that same time only solving them. But the participants were unable to do the experimental task on the basis of instructions alone. They needed training in how to do the explain similarity task using the worked examples and the unguided practice. This made our manipulation potentially stronger because now the experimental group did not just practice comparing whereas the control group did not, they also received training in comparing, which the solve-only control group did not receive. To provide a cleaner and fairer test we also included the solve-with-instructions control group. In the worked examples and guided practice phases, they, like the experimental group, but unlike the solve-only control group, received training in how to compare examples. In the unguided practice, however, they only solved the problems. Fortunately for simplicity of interpretation, the two control groups did not differ, and were collapsed for purposes of analysis. Therefore, the design refers to the training factor as having two, not three, levels.

3 To receive information on the details of the method of scoring, contact the author and ask for the previous, longer version of the manuscript.

4 To receive more detailed information about the errors, contact the author and ask for the previous, longer version of the manuscript.

Appendix A

Examples of Training and Test Problems

Training Problem **-** *Goal: Find Combined Ratio*

A dairy farmer mixed 1 quart of milk that was 2% fat with 3 quarts of milk that was 5% fat. What was the percentage of fat in the whole 4 quarts of milk?

Training Problem **-** *Goal: Find Initial Ratio*

A college student, Bill, and his girl friend, Hillary, attend two different colleges. They have agreed to meet at a location that is exactly half way between them. Bill and Hillary began driving to the meeting place at exactly the same time. Hillary, who always drives at 75 mph. will arrive in 8 hrs. Bill begins by traveling at 80 mph for the first 6 hrs., but he needs to slow down for the last 2 hrs. of the trip because he wants to arrive at the same time as Hillary. At what speed should Bill drive for the last 2 hours?

Test Problem **-** *Old Surface Features, Old Solution Procedure*

A chemist combines 5 qts. of a 40% acid solution with 15 qts. of a 20% acid solution. What is the resulting % of acid of the whole 20 qts. of solution?

Test Problem **-** *New Surface Features, New Solution Procedure*

In an experiment on the effects of smoking on heart rate, the average heart rate for 35 male subjects in the experimental group was 78 beats per minute. The average heart rate for the female subjects in the experimental group was 76. The average heart rate for the entire experimental group was 76.5. How many female subjects were there?