Learning Spurious Correlations instead of Deeper Relations

Norma Chang (nchang@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213 USA

Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213 USA

Marsha C. Lovett (lovett@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213 USA

Abstract

Effective instructional design requires navigating the tradeoff between providing helpful cues to the correct solutions and supplying hints that ultimately detract from what students learn. The present study manipulated the correlations between superficial features and the correct solutions in a set of training problems in the domain of exploratory data analysis and examined their effect on novices with no prior knowledge of statistics. Students who were trained on problems with these spurious correlations performed more poorly on posttest problems lacking these associations, making errors in the direction predicted by the misleading features. The theoretical and educational implications of the outcomes of this practice are discussed.

Introduction

When learning to solve problems based on a new concept, students often are influenced by the surface features of the problems (Chi, Feltovich, & Glaser, 1981; Ross, 1984, 1987; Palmer, 1997). While this reliance on superficial features may facilitate the quick retrieval of previously successful solution methods that are helpful to the current problem, it can also hinder students' abstraction of the deeper concepts that can be transferred to more distant problems. Instructional treatments that draw upon the impact of specific instances, such as learning from worked-out examples (Sweller & Cooper, 1985; Zhu & Simon, 1987) or case studies (Kolodner, 1993), consequently may risk having their effectiveness undermined by students' oversensitivity to surface similarity.

A variety of interventions involving explicit instruction have been examined as possible remedies to alleviate this problem. Clement and his colleagues capitalize on the power of concrete details in advocating the use of "bridging analogies" that gradually develop the fundamental principles from the superficial features by providing a sequence of progressively more abstract analogues to the original "anchor" (Clement, Brown, & Zietsman, 1989; Brown, 1992; Clement, 1993). Other research on analogies underscores the value of the familiar "compare-and-contrast" injunction in highlighting key relationships and conceptual structures within the examples being studied. As described by Gentner and her colleagues, analogical encoding or mutual alignment between simultaneously juxtaposed examples promotes the abstraction and transfer of general principles (Loewenstein, Thompson, & Gentner, 1999; Thompson, Gentner, & Loewenstein, 2000; Loewenstein & Gentner, 2001; Kurtz, Miao, & Gentner, 2001). Similarly, Schwartz and Bransford (1998) endorse the analysis of contrasting cases to boost students' ability to identify the features that differentiate those cases, as well as their comprehension of a subsequent explanation of principles involving those features.

To maximize the potential benefit of instructing the learner to generalize across examples, one should select those examples according to principles that will further highlight the deep concepts and diminish the superficial similarities. The multitude of studies that find minimal transfer when irrelevant features such as the situational context change suggest that participants may be encoding and using these features as cues for their problem-solving strategies (Gick & Holyoak, 1983; Catrambone & Holyoak, 1989; see Barnett & Ceci, 2002, for review). Interpreting these results through an associationist framework yields the implication that varying these features during training would prevent these cues from becoming too strongly learned in the first place. Especially if students may already exhibit predispositions toward noticing and utilizing certain superficial features that they find more salient, instructors should beware of allowing those superficial features to become misleadingly helpful during their problem-solving experiences. Rather, the goal should be to select examples that minimize spurious correlations between irrelevant features and the correct solution method, thereby reducing the noise so that the student has the opportunity to learn the conceptual relationships in the problems. This principle

would apply not just to the situational context, but to the peripheral content in the example problems as well.

Other implications that derive from associative learning theory are that even if students recognize that these spuriously correlated cues are not meaningful, they may still need to inhibit these associations when solving the problems, an especially costly burden under the high cognitive load that accompanies difficult problems. Forming associations with multiple cues may also result in decreased strength for the association to the target cue, if the learning is now spread across many cues. Further, if these superficial features are too predictive of the correct solution strategy, their associations may grow so strong as to block the learning of the association to the relevant feature. To test the applicability of this theory, the experiment described here investigated the impact of such spurious correlations on subsequent problem-solving performance in the domain of exploratory data analysis.

Method

Participants

Eighteen undergraduate students from Carnegie Mellon University were recruited to participate in the experiment. Two participants left the experiment due to visa restrictions on their eligibility for payment. None of the participants had any statistics background beyond high-school mathematics, and all were fluent in English.

Design

The experiment employed a between-subjects design, with participants randomly assigned to condition. Nine students completed the condition incorporating spurious correlations with superficial features in the training problems ("spurious" or "S" condition), and seven students completed the condition in which these features were allowed to vary across representation types ("varied" or "V" condition).

Materials

Problem-Sorting Task Six word problems were written with different combinations of cover story themes (cars, sports, crime) and solution methods (boxplot, scatterplot, contingency table). Each problem was typed on a separate 2.5" x 5.5" card so that participants could easily sort the cards into different groups.

Skills Assessment A paper-and-pencil test consisting of 26 multiple-choice questions was constructed to determine participants' knowledge of relevant statistical definitions and their skills at interpreting and selecting the appropriate type of data display to answer a given question (histogram, boxplot, scatterplot, contingency table).

Data-Analysis Training Problems Sixty problems were written for the training phase of the experiment, thirty per condition. Each problem consisted of a dataset and cover story requiring the student to construct one of the following data representations: pie chart, histogram, side-by-side boxplots, scatterplot, or contingency table. The superficial features manipulated were the cover-story theme, the wording of the question, and the number and combination of variable types presented in the dataset.

In the "S" condition, all problems requiring the same type of data representation were accompanied by a cover story of the same theme, with the question of interest always phrased in the same way. For each representation type, every problem except one was presented with the same number and combination of variable types in the dataset (categorical or quantitative), with the variables presented in the same order. The exception provided three variables so that S-condition participants would gain some experience having to decide which variables were relevant to the problem. Table 1 shows the mappings between each of these features and the representation type. As shown below, all the boxplot problems used the same cover story and wording:

> *Gosset College is looking for patterns in its course evaluations to find ways to improve its introductory classes. Is there a difference in overall course ratings between small classes and large classes?*
>
> *Bonferroni University is performing a grade audit to determine if it has experienced grade inflation in the last ten years. Is there a difference in grades between the class of 1990 and the class of 2000?*

These examples each provided a dataset containing one quantitative and one categorical variable.

Table 1: Superficial Features for Problems in "Spurious" Condition

Representation	Cover Story Theme	Question Wording	Variables in dataset
Pie Chart	demographics	"What percentage…"	1 categorical
Histogram	entertainment	"How would you describe the features of the distribution of…"	1 quantitative
Boxplot	academics	"Is there a difference between… in …"	1 quantitative, 1 categorical
Scatterplot	money	"Is there an association between… and …"	2 quantitative
Contingency Table	health	"Is there a significant effect of … on whether…"	2 categorical

In the "V" condition, the problems were written with the same cover stories and wordings as in the S condition, but varying across all five problem types. No more than two problems of the same type used the same cover story, and every problem of the same type contained a different wording. Half of the problems in the V condition provided two variables and half provided three variables. As shown, problems of the same type (e.g., boxplot) used different cover stories, question wordings, and variable types:

The city of Farrsburgh is holding a referendum to determine how its residents feel about ending welfare. The data table shows how a sample of residents voted, along with their annual income. Is there an association between residents' income and how they voted on the referendum?

A health club claims that its exercise regimen will lead to rapid weight loss. Half of its members are randomly selected to enroll in this program, while the rest spend the same amount of time on the treadmill. The data table shows each member's form of exercise, gender, and percent of body weight lost. Did the exercise-program participants lose a greater percentage of their original weight than those on the treadmill?

Data-Analysis Post-Test Problems The post-test problems were designed with differing degrees of correspondence to the spurious correlations introduced during training. The twenty-five data-analysis problems were divided approximately evenly across the five representation types and across three levels of correspondence (having 0, 1, or 3 features matching the spurious correlations that were incorporated into the S-condition's training problems). The values of the non-matching features were selected to counter each other's influence so that it would be possible to compare their relative effects in misleading students toward different wrong answers. The example below shows a problem with 0 matching features, since the correct representation type for solving the problem is a pie chart, but the cover story, question wording, and variable types are associated with different wrong answers from the S-condition training (contingency table, histogram, and boxplot, respectively):

A study on people's response to sleep deprivation measures how long they sleep before first awakening as an assessment of how deeply they were sleeping. How would you describe the features of the distribution of participants' stages of sleep when first awakening?

This problem was presented with a dataset containing one quantitative and one categorical variable. Note that such a problem would be confusing to the S-participants if they learned the superficial associations presented during training, but the V-participants should not show this effect. None of the posttest questions used wordings that matched the representation type as presented in the V condition.

Procedure

The study began by collecting baseline data on students' initial performance in introductory statistics. The first measure, the problem-sorting task, borrowed from the categorization paradigm employed by Chi, Feltovich, & Glaser (1981) to assess participants' relative use of surface or deep principles in categorizing basic data-analysis problems. Students were given six cards, each with a different problem description on it, and were asked to put the cards into groups in any way they wanted. The second measure, the skills assessment, was a paper-and-pencil pre-test seeking to assess the extent of students' background knowledge of basic concepts in statistics, specifically their abilities to select and interpret the data representations being investigated here.

Students subsequently underwent four days of training, on each day of which they watched a videotaped lecture of a statistics professor at Carnegie Mellon explaining the relevant principles first, and then worked through a portion of the training problems to practice applying those principles. The lectures ensured that all students received explicit instruction on the correct procedures to use in their problem-solving prior to practicing specific examples. Their schedule of practice incorporated opportunities to review what they had learned on previous days so that they could not merely rely on temporal information to determine which representation type to use, since each day introduced only one or two new representation types. Each problem included detailed instructions guiding them through the process of performing the relevant analyses using the statistical software package Minitab after they had thought about which analysis to perform, as well as questions requiring the students to interpret the results they produced. At the conclusion of each problem, students received and reviewed the solution to the problem as a means of providing controlled feedback containing the correct answer.

On the fifth day, students started by completing the same two baseline assessments (problem-sorting task and skills assessment) as they had on the first day. After that, they proceeded to solve the 25 data analysis problems described above as their post-test, without receiving any feedback as to the correct solution. Each day's session lasted no more than 3 hours.

Results

Training Time The number of hours that participants devoted to the training did not differ by condition (M_S = 8.41, SD_S = .848; M_V = 8.27, SD_V = .624; p = .744).

Problem-Sorting Task Participants' responses were analyzed by counting the number of pairs of problems with the same cover story theme (superficial pairings) and the number of pairs of problems that would be solved by the same representation type (deep pairings) that were

categorized in the same group. Analysis of the participants' change in pairing scores showed that both groups learned the relevant principles for solving these problems during the training, in that they utilized more deep pairings and fewer superficial pairings to categorize the problems at posttest, as shown in Figure 1. There was no significant difference between participants in the S and V conditions in their amount of improvement, suggesting that both groups had successfully learned not to use the cover story theme to identify the representation type. Because the problem-sorting task focused on only this feature, assessing the impact of the question wording and variable types requires examining the results from the data-analysis posttest problems, described in a later section.

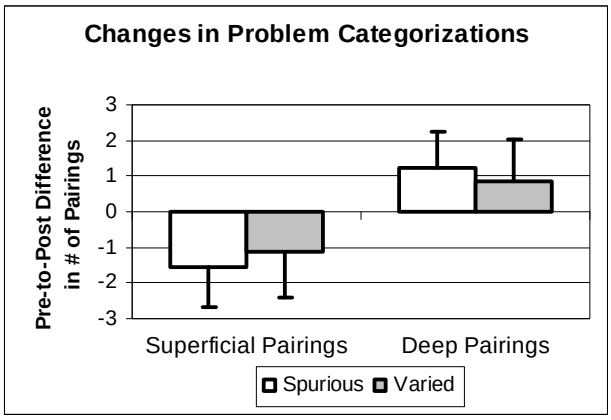

Figure 1: Changes in participants' groupings of word problems from pretest to posttest.

Skills Assessment Participants in both the S and V conditions demonstrated posttest gains significantly greater than zero by single-sample *t*-tests (M_S = 7.33, SD_S = 3.32, N_S = 9, $p_S < .001$; M_V = 9.71, SD_V = 1.89, N_V = 7, $p_V < .001$), as shown in Figure 2. These results further support the claim that the participants learned from the instruction provided during the training. Although V-condition participants did show a larger increase on posttest, this difference between conditions was only marginally significant according to a one-tailed independent-samples *t*-test (p = .057).

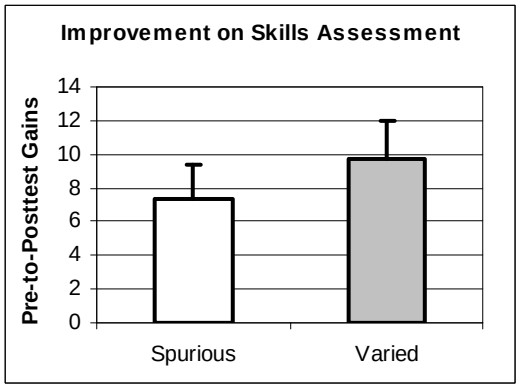

Figure 2: Improvement in test scores after training.

Data-Analysis Post-Test Problems Analyzing participants' overall accuracy in selecting the appropriate graph on the data-analysis problems at posttest shows that V-condition participants[1] performed better than S-condition participants, a difference that is significant when their skills-assessment pretest scores are included as a covariate (M_S = 20.00, SD_S = 2.291, N_S = 9; M_V = 21.33, SD_V = 1.633, N_V = 6; p = .033). Comparing performance between particular types of problems on the posttest reveals that these differences conform to the predicted interaction. For S participants, problems in which all of the superficial features match what they practiced should be easier due to the associations they have learned with these irrelevant features, whereas the V participants should not show the same sensitivity. The results of a 2 (condition) × 2 (number of matching features) ANCOVA on graph-selection accuracy, entering pretest scores on the skills assessment as a covariate, indicate a significant condition-by-match interaction that supports this claim ($F_{1,12}$ = 13.975, MSE = .005, p = .003), as shown in Figure 3. For problems with zero matching features, the S-participants' percentages of correctly-answered questions (M = .728, SD = .168) are significantly lower ($F_{1,12}$ = 13.869, MSE = .006, p = .003) than the V-participants' scores (M = .870, SD = .109). Further, a paired-samples *t*-test shows that the S participants perform significantly worse (t = 4.859, p = .001) on these problems than on problems where all the irrelevant features match what they practiced during training (M = 1.000, SD = .000). In contrast, the performance difference between these problems is not significant for the V participants. Taken together, these findings demonstrate that the spurious correlations presented during the S-condition's training hindered their performance on posttest problems lacking these spurious cues.

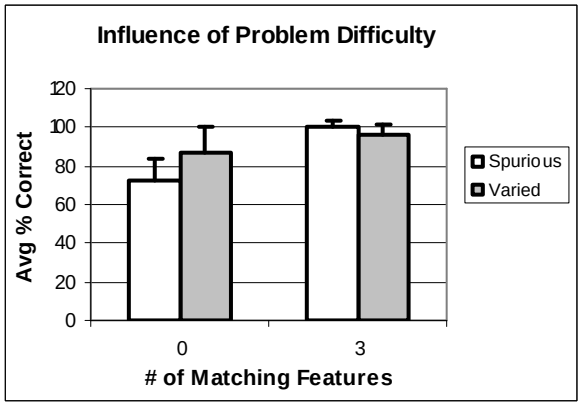

Figure 3: Relative helpfulness of matching features for the spurious condition.

Inspecting participants' error patterns on these problems with mismatched features reveals the impact of the

[1] Data for this task for one participant in the V condition were lost upon transferring computer files.

manipulation still more persuasively. If S-condition participants were indeed influenced by these superficial features when selecting their answers, a disproportionate number of their errors should correspond to the answers prompted by those misleading features. To obtain a relative measure of each participant's susceptibility to the inappropriate use of a particular irrelevant feature (i.e., cover story, wording, or variable types), the number of incidences of such errors made was divided by the total number of opportunities to make that error. Figure 4 shows that errors associated with the variable types predominate in both conditions, and that errors associated with the question wording are much more prevalent in the S condition than in the V condition. Conducting a 2 (condition) × 2 (error type) ANOVA on these error scores comparing wording-based errors against other errors results in a significant interaction ($F_{1,13}$ = 6.165, *MSE* = .002, p = .027). This result is further supported by a chi-square analysis comparing the counts of wording-based errors to other errors for the S condition ($\chi^2(1)$ = 11.75, p = .0006); the same analysis for the V condition is not significant ($\chi^2(1)$ = 0.29, p = .59).

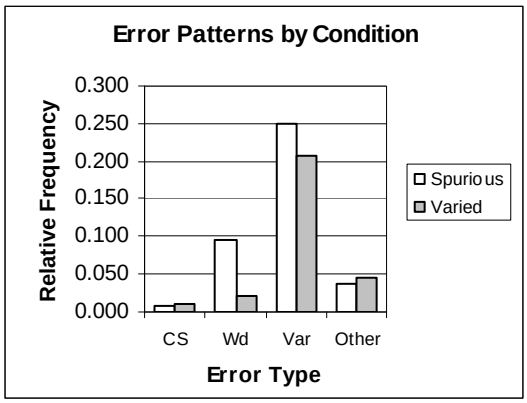

Figure 4: Comparison of errors committed.

Discussion

The results of the analyses of participants' error patterns indicate that training them on problems with spurious correlations with certain superficial features did have the predicted effect of influencing the solutions they chose. Spurious-condition participants learned these superficial associations and utilized them to arrive at incorrect answers at posttest, despite having received explicit instruction on the relevant problem-solving principles prior to practicing example problems during the training.

Of the irrelevant features whose associations were manipulated in this study, the question wordings showed the greatest differential impact on the participants in the spurious condition, while the variable types influenced performance in both conditions and the cover story affected neither. It may seem surprising that the participants were not misled by the cover story themes in solving these problems at posttest, since they had initially used the cover story to categorize the word problems at the beginning of the study. This suggests that by the end of the weeklong training, these students already progressed past the level of novice with respect to their reliance on cover-story information, in contrast to the students described by Chi, Feltovich, & Glaser (1981). One possible reason may be that the cover story theme is not considered to be as meaningful in the domain of exploratory data analysis as is the comparable information in physics (such as the presence of pulleys, inclined planes, etc.), and thus is easier to learn to reject. Such a view is encouraging in implying that students are not equally influenced by all superficial features, but are capable of distinguishing at some level which features are not meaningful and disregarding this information. Another possibility may be that the problem-sorting task itself called participants' attention to the irrelevance of the cover story, since they may have expected that they should sort the problems differently upon posttest, and no other features had been manipulated in this task.

In contrast to their ability to reject cover-story information, these study participants exhibited a striking susceptibility to spurious correlations with the question wording, following these cues even when other problem features should have led them toward different answers. This result gives rise to especially cautionary implications, since teachers sometimes deliberately instruct students to seek out these verbal cues to help them decide upon a solution method. As shown by their comparably good posttest performance on the problem-sorting task and skills assessment, students in the spurious condition successfully learned what information they should use to decide which display type was appropriate, yet they still fell prey to the trap set by the question wording when solving these problems. These results have since been replicated in more recent research using the same paradigm with a larger sample of participants (Chang, Koedinger, & Lovett, in preparation).

More difficult to eradicate is the most common error that both groups demonstrated, that of using the variable types presented to decide which type of display to create. It is possible that participants may have learned to rely on the variable types in the dataset to reach the correct answer even without having practiced any problems with these spurious correlations. As was explained in the explicit instruction provided via the lectures, the correct solution method was to determine the number and type (categorical vs. quantitative) of variables relevant to the problem, and then choose the appropriate data representation. Students may have learned the latter part of this rule correctly but failed to apply the first part of it reliably, namely, identifying which variables were relevant.

Even though the V-condition participants had practiced selecting which of the two or three variables present were actually relevant to the problem, these students nevertheless seemed to be readily tempted by this information on the posttest problems. It may be that this step of selecting the

relevant information needs to be practiced still more frequently or trained even more explicitly for students to execute it reliably. This may be especially likely since the problems that students solve in artificial classroom contexts tend to package information so cleanly that students develop the habit of assuming that everything included in the problem is relevant. Given that students have learned these strong biases toward using all the information in a problem, designing problems without spurious correlations becomes even more critical, both to free them of this habit and to reduce the likelihood of latching onto features that may be present but not deeply meaningful.

The prevalence of committing this error even in the V condition suggests that the extra practice they received in identifying the relevant variables still was not sufficient to protect them against this error. This argues against the possible explanation that deeper processing due to solving longer problems during training may be responsible for their superior performance. Moreover, there was no significant difference in training time between the two conditions. Nevertheless, it may be helpful to explore this phenomenon further using problems that are more similar in complexity to assess the impact of the spurious correlations alone.

The implications of these findings are that the common practice of simplifying and scaffolding the learning process by providing too many helpful cues deserves to be examined more closely. If the same mechanism that facilitates performance in the short term actually inhibits deeper learning and transfer in the long term, we should beware of offering this crutch too readily in our attempts to boost problem-solving performance. Where possible, students need exposure to problems in which the irrelevant features vary sufficiently to convincingly demonstrate their irrelevance, even if this variability is introduced gradually. Designing problems without these spurious correlations removes the temptation to rely on predictive but superficial cues, while also exemplifying the range over which the features may vary and the diverse contexts to which the principles apply, thereby allowing both students and their teachers to focus on developing a richer understanding of the deeper conceptual relations instead.

References

Barnett, S.M., & Ceci., S.J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637.

Brown, D.E. (1992). Using examples and analogies to remediate misconceptions in physics: Factors influencing conceptual change. *Journal of Research in Science Teaching, 29*(1), 17-34.

Catrambone, R., & Holyoak, K.J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1147-1156.

Chang, N., Koedinger, K.R., & Lovett, M.C. (2003). *What students learn from spurious correlations*. Manuscript in preparation, Carnegie Mellon University.

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.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of Research in Science Teaching, 30*(10), 1241-1257.

Clement, J., Brown, D.E., & Zietsman, A. (1989). Not all preconceptions are misconceptions: Finding 'anchoring conceptions' for grounding instruction on students' intuitions. *International Journal of Science Education, 11*, 554-565.

Kolodner, J.L. (1993). *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann.

Kolodner, J.L. (1997). Educational implications of analogy: A view from case-based reasoning. *American Psychologist, 52*(1), 57-66.

Kurtz, K.J., Miao, C.H., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10*(4), 417-446.

Loewenstein, J., & Gentner, D. (2001). Spatial mapping in preschoolers: Close comparisons facilitate far mappings. *Journal of Cognition and Development, 2*(2), 189-219.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*(4), 586-597.

Palmer, D. (1997). The effect of context on students' reasoning about forces. *International Journal of Science Education, 19*(6), 681-696.

Ross, B.H. (1984). Remindings and their effect in learning a cognitive skill. *Cognitive Psychology, 16*, 371-416.

Ross, B.H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 629-639.

Schwartz, D.L., & Bransford, J.D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475-522.

Sweller, J., & Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59-89.

Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organizational Behavior and Human Decision Processes, 82*(1), 60-75.

Zhu, X., & Simon, H.A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*(3), 137-166.