Running Head: REPRESENTATION OF PRINCIPLES

REPRESENTATION OF ARITHMETIC PRINCIPLES BY NOVICES: KNOWLEDGE OR ATTENTION?

Aaron S. Yarlas

and

Vladimir M. Sloutsky

The Ohio State University

April 4, 2000

PAPER IN PROGRESS

Address correspondence to:

Vladimir M. Sloutsky
21 Page Hall
1810 College Avenue
The Ohio State University
Columbus, OH 43210
Phone: (614) 688-5856
Fax: (614) 292-0321
Email: sloutsky.1@osu.edu

Abstract

Four experiments examined the nature of novice and expert representations for both surface features and principled (relational) properties of arithmetic equations. Experiment 1 used a forced-choice categorization task, in which surface features of equations (e.g., numbers) competed with principled properties of mathematics (associativity and commutativity). It was found that experts were more likely to focus on principles in their judgments than were novices, who focused more often on surface features. Experiment 2, using a similar task, introduced trials in which only principled properties varied. Novices were able to focus on principled properties in this case, but failed to transfer these representations when surface features were reintroduced. These findings indicate that novices had knowledge of the principles, but that they did not attend to them when competing surface features were present. Experiment 3 used a recognition task to determine whether or not novices encode deep principled properties, whereas Experiment 4 used the recognition task to determine the manner in which features and principles are encoded. A process model describing probable mechanisms for novice representations of principled (relational) properties is discussed.

REPRESENTATION OF ARITHMETIC PRINCIPLES BY NOVICES: KNOWLEDGE OR ATTENTION?

It has been well established that in different knowledge domains (e.g., physics, mathematics, or chess), experts approach problems in a manner different from that of novices (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Larkin, 1983; Simon & Simon, 1978; Reed, Ackinclose, & Voss, 1990). In particular, experts and novices often diverge as to what components of a problem they represent. While experts are more likely to focus on hidden relational properties of a problem, such as equations that can be used to solve the problem, novices are more likely to focus on less important surface features of a problem, such as the problem's content.

Note that although the term hidden relational property refers to properties that can not be directly observed, the term has been used in at least three ways. First, hidden (or "deep") properties are those that require knowledge, such as taxonomic or functional properties (e.g., the fact that a whale is a mammal is hidden). Second, hidden properties are relational as opposed to object properties. For example, given two pictoral dyads, (1) daughter-mother and (2) mother-grandmother, the mother in the first picture corresponds to the mother in the second picture in terms of object properties (it is the same object), whereas it corresponds to the grandmother in terms of hidden relational properties (it is the same relation: the mother of *X*). Finally, hidden properties constitute important principles in a knowledge domain (e.g., Boyle's law) as opposed to surface or accidental properties (e.g., the story line of a particular problem). It seems that the three meanings are interdependent as they jointly emphasize that hidden properties are (1) knowledge-dependent, (2) not directly perceptible, and (3) relational. In this paper, we will use the terms deep relational features or deep principles as referring to those hidden properties that have (1), (2), and (3).

The term surface feature also affords several meanings. For example, clinical cases could be grouped on the bases of medical symptoms (e.g., high fever vs. headache), demographic variables (e.g., age and gender), or patients' hair colors. All of these would be surface features as opposed to deep relational properties, such as clinical syndromes (e.g., bronchitis or depression) relied upon by experts (see Custers, Boshuizen, & Schmidt, 1998, for a discussion). However, medical symptoms and demographic variables could be "meaningful" surface features, because they could be predictive of a clinical syndrome, whereas patent's hair color is a superfluous surface feature. Although the distinction between meaningful and superfluous surface features may appear arbitrary, it has proved useful in several knowledge domains (see Reeves & Weisberg, 1994 for a discussion). In the current studies, we used those surface features that were superfluous with respect to the deep relational properties.

There is a large body of literature indicating that in problem solving, reasoning, learning and transfer, and problem categorization, novices tend to focus on surface features rather than on deep relational properties. These effects have been demonstrated in a variety of knowledge domains. These domains include chess (Chase & Simon, 1973), mathematics (Blessing & Ross, 1996; Hinsley, Hayes, & Simon, 1977; Schoenfeld & Herman, 1982; Bassok, 1996, 1997; Novick, 1988; Reed, et al, 1990; Ross & Kilbane, 1997; Silver, 1981), physics (Chi, et al 1981; Simon & Simon, 1978; Larkin, 1983; Larkin, McDermot, Simon, & Simon, 1980), and computer programming (Adelson, 1984). Similar effects have been observed in a variety of knowledge-lean domains, such as deductive and inductive inference. When presented with deduction problems, untrained reasoners often tended to ignore the argument's logic (i.e., its deep structure) while relying on the argument's surface features, such as content and believability (Cheng & Holyoak, 1985; Evans, Newstead, & Byrne, 1993; Johnson-Laird & Byrne, 1991; Byrne, 1989). When presented with induction and analogy problems, novices often ignored deep relational structure while relying on the surface features (Gentner, 1989; Gentner & Toupin, 1986; Kotovsky & Gentner, 1996; Holyoak & Koh, 1987).

While there is little disagreement that novices focus on surface features, several issues remain unclear. Why do novices tend to represent surface features and not deep relational properties? And why does instruction often fail in helping novices represent deep relational properties, such as rules of logic (Nisbett, Fong, Lehman, & Cheng, 1987) or algebraic equations (Reed, et al 1990; although see Schoenfeld & Herrmann, 1982, for diverging results). In addressing these issues, we will also touch upon another issue of whether superfluous surface features constitute an impediment or aid in extracting the deep relational structure of a task.

One possible explanation of novices' tendency to focus on surface features is that novices merely have little knowledge of deep structural relations. However, while this possibility is capable of explaining expert-novice differences in extremely knowledge-demanding domains, such as medical diagnostics, chess, or advanced physics, it falls short of explaining these differences in fairly simple domains, such as elementary mathematics and physics. For example, researchers examining novices' representations in mathematics and physics often drew examples from students' textbooks, thus reasonably assuming that students should be familiar with the deep structure underlying these problems (Chi, et al, 1981; Larkin, 1983; Novick, 1988). The credibility of the lack of knowledge explanation is further undermined by findings that even those novices who receive instruction in a domain often continue to focus on surface features rather than the deep structure of a problem. These has been demonstrated across a variety of knowledge domains, including mathematics (Morris & Sloutsky, 1998), logic (Nisbett, et al. 1987), and physics (Kaiser, McCloskey, & Proffitt, 1986; McClosskey, 1983). Finally, the fact that findings on novices' representations in knowledge-lean domains are compatible with those in knowledge-rich domains (Cheng & Holyoak, 1985; Evans, Newstead, & Byrne, 1993; Johnson-Laird & Byrne, 1991; Byrne, 1989, Gentner, 1989; Gentner & Toupin, 1986; Kotovsky & Gentner, 1996; Holyoak & Koh, 1987) makes the low knowledge hypothesis even less plausible.

Another possibility is attentional: even when novices know about deep relations and are capable of extracting these relations, they still fail to represent these relations because surface features are more prominently present in the problem. In failing to represent relational properties, they may either fail to encode these relations, or they may initially encode these relational properties, but discard them in favor of more salient surface features. It is also possible that novices use surface features more frequently than relational properties, and, as a result, when surface features are well-known, they are more accessible than deep relational properties (cf. Anderson, 1990).

The salience of surface features could be demonstrated using task presented in Figure 1. If one considers similarity of objects, then Test 2 is more similar to the Target (both comprise circles), whereas if one considers similarity of relations, then Test 1 is more similar to the Target (both comprise a monotonic increase). Note that it is easier to overlook relational similarity than to overlook object similarity: while the latter is directly perceptible, the former is not. These intuitions have been corroborated in a series of studies where participants often failed to represent relational similarity in tasks similar to those depicted in Figure 1 (Gentner & Toupin, 1986; Kotovsky & Gentner, 1986; Gentner & Markman, 1997). As mentioned above, the attentional explanation of the tendency to represent surface rather than deep relational properties comprises several options: novices may fail to encode deep relational properties, or they may initially encode both surface and deep relational properties but discard the latter in favor of more salient surface features.

The first option, of course, makes surface features more benign than the second, because these features are often predictive of the deep relational structure (e.g., Bassok, 1996; 1997; Blessing & Ross, 1996). Therefore, if novices would fail to encode deep relational properties anyway, attention to the surface structure may facilitate novices' learning, transfer, and problem solving (see Bassok, 1997, for a discussion). The second option suggests that surface features, by their virtue of being salient, interfere with the process of representation of deep relational properties (Bassok, 1997; Gentner, 1989; Gentner & Toupin, 1986). Note that because in the current experiments, we did not use surface features that were predictive of deep relational properties, answers provided by the current research would be limited only to superfluous surface features.

To distinguish among the above-mentioned explanations (i.e., knowledge vs. attention), we constructed tasks where surface features were completely superfluous and were not predictive of deep relational properties. Do novices know the deep relational properties? Do they represent both surface features and deep relational properties? And do they encode both surface features and deep relational features? The present research addresses these questions.

Of course, it is also possible that experts and novices differ dramatically with respect to their general intelligence levels, with experts being a self-selected group, while novices represent more general segments of the population. At the same time, different intelligence theorists agree that the ability to detect relations is indicative of general ability (Horn & Noll, 1994; Kyllonen, 1994; Sternberg, 1985). While this possibility is not likely to tell the whole story (e.g., Schoenfeld & Herrmann, 1982, demonstrated that novices can improve their problem representation), it might tell a part of the story. Therefore, we deemed it necessary to control for overall ability levels when comparing experts and novices.

The goal of the current studies is to provide a detailed examination of problem representation in novices along the lines of the questions outlined above. To achieve this goal, we deemed it necessary to control for the intelligence and knowledge factors, while manipulating representational factors. In controlling for knowledge factors, we (a) used simplified tasks and (2) selected only those deep relational properties (or principles) that were well familiar to novices in mathematics. In particular, we selected the commutative and associative properties of arithmetic, because these principles are likely to be familiar to the majority of college undergraduates (Everyday Mathematics: Teacher's Reference Manual, 1998).

In this paper, we present four experiments. In Experiment 1, experts and novices in mathematics were asked to group arithmetic equations. These groupings could be based either on the commonality of surface features (e.g., numbers used, the number of constituent elements in the equations) or on the commonality of a deep mathematical relation (principles of commutativity or associativity). It has been argued that researchers often arbitrarily categorize certain relations as "deep and structural" (Brown, 1989), taking their intuitions as face validity. To counter this argument, we deemed it important to ascertain that associativity and commutativity are treated as important mathematical relations not only by math textbooks, but also by mathematics experts.

In Experiment 2, we introduced a two-phase grouping task. During the first phase, deep relations were "unmasked," such that surface features were not varied among the compared equations. During the second phase, the deep relations were "masked" again by reintroducing competing surface features. In this experiment, we were particularly interested in performance in the "unmasked" phase and in transfer from the "unmasked" to the "masked" phase.

In Experiment 3, equations similar to those used in Experiments 1 and 2 were presented in a “new/old” recognition paradigm. The goal of this experiment was to examine whether or not novices encode deep relational properties. This experiment was designed to help distinguish between the two options mentioned above: (a) that novices fail to encode deep relational properties, or (2) they initially encode both surface features and relational properties, but discard the latter in favor of more salient surface features.

Experiment 4 used the same methodological paradigm as Experiment 3, but decreased the encoding time in which participants were exposed to the equations in the study phase. The goal of Experiment 4 was to examine the manner in which surface and relational properties are encoded. For example, if surface and relational properties are encoded in a parallel manner, then shortening the encoding time should affect recognition on both of these features. However, if the are encoded in a serial manner, then it is more likely that recognition for only one type of feature will be effected.

EXPERIMENT 1

Method

Participants

Three samples, including two novice groups and one expert group, were selected for the current experiment. The first group of participants consisted of 25 undergraduates in an introductory psychology course at a large Midwestern university who participated for course credit. This group had an average age of 19.78 years (SD = 1.38), with 11 women and 14 men.

This group of mathematics “novices” was contrasted with a group of mathematics "experts". The expert group consisted of 20 graduate students in a Mathematics department at the same university who participated for payment of ten dollars. This group had an average age of 28.88 years (SD = 6.05), with 7 women and 13 men.

Differences between the "expert" and "novice" groups were not limited to expertise. Experts were also older and they might represent a self-selected group with respect to an overall ability. Therefore, we deemed it necessary to select a matching group that would be similar to experts in terms of age and overall ability, while differing in the level of expertise. This matching group consisted of 16 graduate students in a History department at the same university who participated for a payment of ten dollars. This group had an average age of 29.93 years (SD = 4.67), with 8 women and 8 men.

Overall scores on the GRE were used to assess whether the two groups of graduate students (History and Mathematics) differed on ability level. The average reported composite scores for the two groups were similar (2029 for Mathematics students vs. 2017 for History students; $t < 1$), indicating that the two samples were similarly matched on general ability, and differed mainly in domain-specific expertise.

Materials

Five features of arithmetic equations were used in Experiment 1. Two of these features were considered "principled properties", in that they represented deep, relational principles of mathematical operations: the associativity and commutativity principles. The former states that for addition, subtraction, and multiplication, constituent parts can be decomposed and recombined in different ways (e.g., $a + b = [a – c + c] + b$). The latter states that the order of elements is irrelevant for addition and multiplication (e.g., $a + b = b + a$). The other three features were nonprincipled surface features that occur in arithmetic equations: (1) numbers (e.g., 6, 14); (2) sign (e.g., -, +); and (3) the number of constituent terms in an equation. The numerical solutions of equations were controlled for by making these solutions either all equal or all different for each trial.

A forced-choice similarity paradigm was used in this experiment. Participants were presented with three cards at a time, a target card and two test cards, each which had printed on it an arithmetic equation. Participants were instructed to match the problem on the target card to one of the test problems with which they believed it was most similar. Each of the two test problems shared one feature with the target problem, and differed on the feature that the target shared with the other test problem, with all other features held constant. All five features were pitted directly against each other, with the exception of the two principled properties, yielding a total of nine feature comparisons. These nine comparisons, along with their respective problem templates, are presented in Table 1.

There were four exemplar arithmetic equations representing each of the nine comparison sets, resulting in a total of 36 trials presented to participants. The numbers used in the arithmetic equations ranged from 1 to 14, and the operations used included addition, subtraction, and multiplication.

Procedure

Participants were taken individually by a male experimenter into a small, quiet room. All subjects were given the following cover story. They were told that a math professor (and thus mathematics expert) at the university had previously sorted cards with arithmetic equations on them on the basis of which problems the professor had judged to be more similar to each other, but that the experimenter had lost the results of this sorting. The participant was then instructed to match the equations together like the professor had, based on which equations were most similar to each other. The gender of the professor was never identified.

Participants received 9 practice trials, which included one exemplar for each of the feature comparisons presented in Table 1. These practice trials were similar to the experimental trials. These trials were used to acquaint participants with the experimental task, and to avoid random responding in initial trials.

Three trials for each of the nine feature comparisons presented in Table 1 resulted in a total of 27 experimental trials, which took approximately 25 minutes. Trial order was determined using a block randomization procedure, such that trials were randomly ordered without replacement within each of the three blocks of the nine comparison-types. The positioning of the test items in relation to the target (i.e., left or right) was counterbalanced across feature comparison type. In addition to recording which test problems were selected by the participants, their verbal explanations for their similarity judgments were also recorded by the experimenter.

Results and Discussion

One of the goals of this experiment was to examine participants' knowledge of principles in question. To achieve this goal, we considered as choices indicating knowledge only those for which the participants' explanation of the choice was consistent with the principle. This was done because participants could select principled test stimuli for a reason that might have nothing to do with the principle in question. Only explanations directly referring to the principle in questions were considered choice-consistent. For example, when choosing a commutativity target, an explanation such as “In both equations, they have the same numbers on both sides, just in different orders” was considered choice-consistent, while an explanation such as “Both equations have a ‘7’ in them” was considered choice-inconsistent. The advantage of this measure is that even a single principled choice accompanied by an explanation indicates knowledge of a principle in question. The proportion of consistent choices for each principle across the 18 comparisons in which features were involved is the dependent variable used in the forthcoming analyses. We also present proportions of participants who had at least one trial for which they focused on a principle and provided a principle-consistent explanation.

An analysis of novices’ choices and explanations, across both principled and surface features, yielded a high level of systematicity between the two. In other words, novices’ explanations were highly consistent with the Test equation chosen across all five features (95% for undergraduates, and 97% for History graduate students). This trend was also observed for the two principled properties: when choosing a principled Test equation, 79% of undergraduates and 90% of History graduate students gave an explanation consistent with the principle in question. This indicates that novice participants were able to externally verbalize the reasons for their choices, including for principles.

The degree to which novices made explanation-consistent choices for each of the five features was examined using Chi-square analyses, followed by analyses of standardized residuals for each of the novice samples. For both samples, the omnibus Chi-squares were significant (both $\chi^2 > 70$, ps $< .001$), enabling for a more detailed analysis of the residuals. For undergraduates, this analysis yielded that explanation-consistent choices based on number of elements were most frequent, followed by in descending order by choices based on sign, numbers, commutativity, and associativity (all zs > 2.4, all ps $< .01$). Patterns of standardized residuals also indicated that History graduates made explanation-consistent choices for number of elements and sign equally frequently, less frequently for number and commutativity, and least frequently for associativity.

The degree to which participants in each sample made explanation-consistent principled choices was analyzed using a one-way Analysis of Variance (ANOVA) for each principle across the three samples. The ANOVA for explanation-consistent associativity choices yielded a significant difference among the samples in the proportion of choices made, F (2, 58) = 25.35, MSE = .11, $p < .001$. Bonferroni post-hoc tests indicated that there was no significant difference ($p > .4$) in the percentages of explanation-consistent associativity choices between the undergraduates (M = 10.66%, SD = 26.15%) and history graduates (M = 26.39%, SD = 38.89%). However, the expert sample gave more explanation-consistent associativity choices than either of the novice samples (M = 80.00%, SD = 44.78%), both ps < .001.

The analysis of commutativity judgments yielded a similar pattern of results. The ANOVA for explanation-consistent commutativity choices also indicated that there was a significant difference among the samples in the proportion of choices made, F (2, 58) = 16.11, MSE = .11, $p < .001$. Bonferroni post-hoc tests indicated that there was no significant difference ($p > .4$) in the percentages of explanation-consistent commutativity choices between the undergraduates (M = 20.44%, SD = 30.71%) and history graduates (M = 36.81%, SD = 35.54%), and the expert sample gave more explanation-consistent commutativity choices than either of the novice samples (M = 77.22%, SD = 35.95), both ps < .005.

In addition to the Bonferroni post-hoc tests, effect sizes (i.e., Cohen d's) were computed for the difference among the three samples for proportion of explanation-consistent principled choices. There was a medium effect size for the difference between History graduate students and undergraduates (d = 0.49), while very large effect sizes were found for the differences between Mathematics graduate students and undergraduates (d = 1.40) and between Mathematics graduate students and History graduate students (d = 1.09). This provides a direct indication that differences due to expertise were considerably larger than those due to age or general ability.

Another important question is whether the use of principles by participants stem from the fact that some novice participants focused on principles occasionally or because few participants used principles consistently. The analysis of individual patterns of responses is presented in Table 2. Data in Table 2 indicate that while the majority of experts consistently focused on principles (75% of participants), undergraduate students and history graduates were mostly split between those who never focused on principles and those who focused on principles occasionally.

Results from Experiment 1 point to several important regularities. First, experts were found to consistently focus on principles when categorizing arithmetic equations, whereas novices were more likely to focus on surface features rather than on principles; even when novices did focus on principles, they most often did so inconsistently. Second, expert-novice differences were not limited to age or general intelligence: history graduate students and math experts, equally aged groups with similar levels of overall intelligence, exhibited large differences in using deep principled properties. Third, novices exhibited consistency between their choices and their explicit explanation for these choices. Therefore, they did not respond haphazardly or randomly to the task; they systematically focused on a particular feature for each trial.

Thus the experiment allows us to validate the structural importance of the principles in question and to undermine the possibility that general ability accounts for expert-novice differences. However, Experiment 1 left an important question unanswered: it remains unknown why many novices failed to focus on deep principled properties, focusing instead on a particular surface feature. One possibility is that many novices merely did not know these principles. Another possibility is that novices know these principles, but they either fail to encode these principles, or encode the principles, but discard them in favor of more salient surface features. The goal of Experiment 2 is to provide a detailed examination of the first possibility.

EXPERIMENT 2

To examine participants' knowledge of principles in question, we deemed it necessary to establish whether or not novices focus on principled properties when these features do not compete with surface features. In the current study, then, participants were given a number of trials in which the target problem shared a principled property with one of the test problems, and shared no unique surface features with the other test problem. We refer to these trials as “unmasked” since principled properties are no longer attentionally “masked” by surface features. If participants know the principled properties used in Experiment 1, then "unmasking" these principled relations via eliminating competing surface features should increase the proportion of principle-based choices.

In addition, in the current experiment the “unmasked” trials are followed by “masked” trials equivalent to the trials in Experiment 1, in which the surface features were reintroduced to compete with principled properties in participants’ similarity judgments. This will enable the examination of the degree to which representations of principled properties will be maintained, or whether the surface features will draw attention away from principled properties, such that there is no transfer of representation due to the positive learning set. If the former is true, then it is expected that participants’ explanation-consistent principled choices will be more frequent for the subsequent “masked” trials than they were in Experiment 1; if the latter is true, then there should be no difference between the frequency of these choices.

Method

Participants

A “novice” sample of 19 undergraduates in an introductory psychology course at a large Midwestern university participated for course credit. This group had an average age of 21.79 years (SD = 6.49), and included 12 women and 7 men.

Materials and Procedure

The same five features of arithmetic equations (i.e., associativity, commutativity, number, sign, and number of elements) used in Experiment 1 were used in Experiment 2. The same nine comparisons used in the previous experiment were again used here for the last 27 trials (three trials for each of the nine comparisons). In addition, in the current experiment, the first eight trials consisted of “unmasked” comparisons, thus leading to a total of 35 trials.

For the “unmasked” trials, each of the two principled properties (i.e., commutativity and associativity) was compared four times against ‘control’ problems. For these trials, the two test problems were equivalently similar to the target on nonprincipled features, while one test problem shared a principled property with the target problem. For example, for an unmasked-commutativity trial, the target equation was 2 + 6 + 8 = 6 + 8 + 2, the commutativity Test equation was 11 + 1 + 5 = 1 + 5 + 11, and the control Test equation was 3 + 10 + 4 = 12 + 4 + 1.

The same procedure used for Experiment 1 was used for the current experiment, with the exception that there were now thirty-five trials (eight “unmasked” trials followed by twenty-seven “masked” trials) rather than thirty-six trials. Because "unmasking" was believed to make principles transparent, practice trials were not used in the current experiment.

Results and Discussion

For the current analyses we particularly focus on all "unmasked" trials and the 18 "masked" trials where surface features were pitted against commutativity and associativity. We first analyze performance in "unmasked" and "masked" trials across the three groups of novices. For purposes of clarity, we will refer to "masked" trials in Experiment 1 as Masked 1, whereas "masked" trials in Experiment 2 will be referred to as Masked 2. Again, when analyzing performance, we will consider only those choices for which the participants' explanation of the choice was consistent with the principle.

Overall percentages of explanation-consistent choices for each principle for Masked 1, Unmasked, and Masked 2 trials are presented in Figure 2. As evidenced in the figure, participants' explanation-consistent principled choices on Unmasked trials were generally greater in comparison to the choices on Masked 1 trials in Experiment 1. Independent-samples t-tests were used to analyze these differences for each principle. The analysis yielded that Undergraduates gave more explanation-consistent commutativity choices for Unmasked trials than for Masked 1 trials ($t = 8.59$, $p < .001$), though there was a marginally significant difference in the amount of explanation-consistent associativity choices ($t = 1.81$, $p = .078$).

The percentage of participants' explanation-consistent principled choices for Unmasked and Masked 2 trials were analyzed for each principle by using paired-samples t-tests. While the percentage of participants' explanation-consistent associativity choices for Unmasked trials (M = 27.63%, SD = 36.22%) were higher for those than for the subsequent Masked 2 trials (M = 18.13%, SD = 27.27%), this difference failed to reach statistical significance, $t(18) = 1.07$, p = .30. The difference between the percentage of participants' explanation-consistent commutativity choices for Unmasked trials (M = 90.79%, SD = 20.77%) and Masked 2 trials (M =35.09%, SD

= 36.71%), however, was statistically significant, $t(18) = 7.23$, $p < .001$. These differences indicate that there was not pure transfer of representations from Unmasked to Masked 2 trials: once principled properties had to compete again with surface features, the number of explanation-consistent principled choices decreased.

An important question is whether the transfer led to a significant increase of explanation-consistent principled choices compared to when participants were never exposed to Unmasked trials. That is, whether being exposed to a positive learning set significantly increased subsequent attention to principles. To answer this question, we compared participants' explanation-consistent principled choices on the Masked 1 and Masked 2 trials. While the proportions of explanation-consistent principled choices are somewhat larger for each principle on Masked 2 trials than for Masked 1 trials, t-tests revealed that neither of these differences are statistically significant (both $ts(42) < 1.4$, both $ps > .15$). Thus, the positive learning set of the Unmasked trials had a small but nonsignificant effect on the degree to which participants represented principled properties of mathematics problems.

Overall, results of Experiment 2 indicate that 100% of the participants exhibited knowledge of principles in question (i.e., provided an explanation-consistent principled choice on at least one trial), focusing on these principles in Unmasked trials. However, once surface features were reintroduced, representation of principled properties attenuated to levels similar to Experiment 1.

These results allow us to eliminate the lack of knowledge hypothesis. Indeed if participants had not known the principles in question, they should have not focused on these principles in Unmasked trials, but most of them did; even for associativity, where effects were smaller than for commutativity, over half of participants provided at least one explanation-consistent principled choice.

Having established that novices know the principles in question, we will examine whether or not results observed in masked trials of Experiments 1 and 2 stem from attentional factors. These factors include participants' failure to encode principled properties, or encoding the principles but discarding them in favor of more salient surface features. Experiment 3 was conducted to distinguish between these two possibilities.

EXPERIMENT 3

The goal of the current experiment was to distinguish between two possibilities underlying novices' failure to represent deep relational properties. Recall that one possibility was that novices fail to encode principled properties, whereas another possibility is that they encode and discard these properties. One way of examining encoding is to use recognition or recall memory tasks that have been frequently used to examine encoding of both adults and young children (e.g., Siegler, 1978). In the current experiment, we used an ‘old/new’ recognition procedure.

The ‘old/new’ recognition procedure affords the creation of a set of foils, such that patterns of false alarms point to what has been encoded and committed to memory and what has been left out. In the study phase, novices were given a set of arithmetic equations. These equations all utilized a principled property, either commutativity or associativity. In addition, these equations all used consistent levels of two surface features: all equations used numbers ranging between 1 and 9, and all used either 5 or 6 numbers in the equation. In the recognition phase of the experiment, in addition to ‘old’ items, four combinations of ‘new’ equations were presented as foils. Half of these foils, which we refer to as ‘feature +’ foils, maintained the same levels of features as used in the learning phase (i.e., numbers ranging between 1 and 9, and either 5 or 6 numbers in the equation), while the other half of the foils, which we refer to as ‘feature -’ foils, violated these categories (i.e., numbers greater than 9, and either 4 or 7 elements in the equation). Also, half of the foils, which we refer to as ‘principle +’ foils, maintained the use of one of the two principled properties, while the other half, which we refer to as ‘principle -’ foils, did not use any principled properties in the equation. The two levels of the two kinds of properties (feature being either + or -, and principles being either + or -) were fully-crossed, thus creating the four combinations of foils: feature+ /principle+ (F+/P+), feature + /principle - (F+/P-), feature-/principle + (F-/P+), and feature-/principle - (F-/P-).

Figure 3 presents possible response patterns in the recognition task. If participants do not encode principled properties (possibility 1 in Figure 3), they should base their responses solely on the presence or absence of surface features. When surface features are absent (F-/P- and F-/P+ foils), participants should produce fast and accurate "New" responses, whereas when surface features are present, participants should produce "Old" answers. On the other hand, if participants encode both surface features and principled properties (possibility 2 in Figure 3), they should respond "Old" when both surface and principled properties are present and they should respond "New" when either surface features or principled properties are absent.

Method

Participants

Twenty-one undergraduates in an introductory psychology course at a large Midwestern university participated for course credit. Participants had an average age of 19.4 years (SD = 2.29 years), with 16 women and 5 men.

Materials and Procedure

All participants were run individually at a personal computer. All materials were presented using Superlab Pro for Windows, version 1.05 (Cedrus Corporation, 1997).

The experiment consisted of three phases: the study phase, the distraction phase, and the recognition phase. In the study phase, participants were presented with thirty arithmetic equations, which they had been instructed to memorize. All thirty equations used addition, used numbers ranging from 1 to 9, contained either 5 or 6 numbers, and used either the associative or commutative principle (half for each). Each equation was centered and presented in black type on a white screen for ten seconds, with a two-second interval, during which only the white background was seen, between each. The order of equations given was randomized.

A distraction phase followed the study phase for the purpose of clearing participants' short-term memory. For the distractor task, participants were presented with ninety letters, for which they had been instructed to indicate whether the letter was a vowel (by pressing the "Z" key on the keyboard) or a consonant (by pressing the "M" key). Each letter was centered and presented in black type on a white screen. This phase took approximately three minutes.

Following the distractor phase was the recognition phase. Participants were instructed that they would be presented with a number of arithmetic equations, some of which had been presented to them earlier and some which had not been presented earlier. They were further instructed to indicate whether each equation had been presented earlier or not, by pressing the "Z" key on the keyboard if the item was presented earlier, or the "M" key if it had not been presented earlier.

There were a total of sixty equations presented in the recognition phase. Each equation was centered and presented in black type on a white screen. The order of equations presented in this phase was randomized. These equations fell into five categories, with twelve exemplars for each category. The first category contained ‘old’ equations, which consisted of equations that had been presented earlier in the learning phase. The remaining four categories were foils, in that they contained new problems that had not been presented in the study phase. The first type of foil consisted of "feature + /principle +" (F+/P+) equations that used similar surface feature as the original equations (i.e., numbers between 1 and 9, and either 5 or 6 numbers), and used either the commutativity or associativity principle (e.g., 7 + 1 + 4 = 1 + 7 + 4 and 9 + 7 = 5 + 4 + 7). The second type of foil consisted of "feature + /principle -" (F+/P-) equations that used similar surface features as the original equations but did not use either the commutativity or associativity principle (e.g. 3 + 1 + 9 = 2 + 5 + 6 and 4 + 9 = 3 + 8 + 2). The third type of foil consisted of "feature - /principle +’ (F-/P+) equations that used surface features different from those used in the original equations (i.e., numbers greater than 9, and either 4 or 7 numbers), and used either the commutativity or associativity principle (e.g., 7 + 12 = 12 + 7 and 5 + 1 + 15 = 3 + 2 + 1 + 15). The fourth type of foil consisted of "feature - /principle - ’ (F-/P-) equations that used surface features different from those used in the original equations and did not use either the commutativity or associativity principle (e.g., 8 + 3 + 12 = 5 + 1 + 4 + 13 and 6 + 11 = 2 + 15).

Results and Discussion

In this section, we will first examine the number of OLD responses given for each of the five foils (i.e., correct acceptance of “Old” targets and incorrect acceptance of all other foils). After that we will specifically compare participants responses (in terms of accuracy and latency) to "Old" targets vs. F+/P- foils.

The percentages of OLD responses across foil types are presented in the left-hand panel of Figure 4. These results indicate that the proportion of OLD responses for F-/P- (M = 0.012, SD = 0.004), F-/P+ (M = 0.016, SD = 0.003), and F-/P+ (M = 0.19, SD = 0.28), were significantly lower than chance (all ts(20) < -4.99, ps< .001), whereas for F+/P+ (M = 0.84, SD = 0.15) and Old Targets (M = 0.88, SD = 0.11), the proportion of OLD responses were significantly greater than chance, both ts(20) > 10.19, ps< . 001. These findings indicate that participants accurately rejected foils when either surface or principled properties were absent, but considered foils as OLD when both surface and principled properties were present. The latter finding is expected because F+/P+ foils were categorically indistinguishable from Old Targets, since both surface features and principled properties present in Old Targets were also present in F+/P+ foils. These results indicate that participants took the task seriously and were providing rather accurate responses. These results also eliminate the first possibility presented in Figure 3, while supporting the second possibility. Indeed, participants rejected as NEW those foils where either surface features or principled properties were absent (i.e., F-/P+, F+/P-, and F-/P-), while accepting as OLD only those foils where both surface features and principled properties were present.

A one-way repeated measures ANOVA points to significant differences in accuracy among foils (F (4, 80) = 96.9, MSE = .003, p < .001. Post-hoc comparisons with Bonferroni adjustments indicate that F-/P- and F-/P+ (correct answer "New") were most accurate (all ts(20) > 3, all ps < .05), whereas F+/P+ foils were least accurate (all ts(20) < -7.5, all ps < .001). At the same time, there was no significant differences between F+/P- and Old Targets, t (20)= 1.1, p > .9. Another important finding is that participants equally frequently false alarmed to both commutativity (101 out of 126 responses) and associativity (105 out of 126 responses) F+/P+ foils, $\chi^2 < 1$, p > .5. This latter finding, in conjunction with accurate rejections of F+/P- foils, suggests that participants equally well noticed and encoded both principles of commutativity and associativity. It is quite important because while it could be argued that commutativity has a perceptual component, associativity is a completely non-perceptual principle.

The right-hand side panel of Figure 4 presents latencies to correct responses across F-/P- (M = 1299 ms, SD = 894 ms), F-/P+ (M = 1374 ms, SD = 1013 ms), F+/P- (M = 2985 ms, SD = 1463 ms), and OLD foils (M = 2194 ms, SD = 1507 ms), and all responses for F+/P+ foils (M = 2383 ms, SD = 1315 ms). A one-way repeated measures ANOVA indicates significant differences among the foils, $F(4, 740) = 86.9$, $p < .001$. Post-hoc comparisons with Bonferroni adjustments were used to examine differences in latencies among foils. These tests revealed that F+/P- latencies were significantly higher than those for Old Targets, $t(187) = 5.1$, $p < .001$, and for F+/P+ foils, $t(215) = 10.3$, $p < .001$, but that there was no significant difference between Old targets and F+/P+ foils, $t(217) = 2.2$, $p = .3$. These tests also indicated that latencies of F-/P- and F-/P+ foils were significantly lower than those of the other three foils, all ts < -7.8, all ps < .001. The fact that F+/P- foils elicited the greatest latencies is particularly interesting. The correct response "NEW" to this foil elicited greater latencies than the correct response "NEW" to F-/P- and F-/P+ foils, and greater latencies than the correct response "OLD" to Old Targets. This delayed responding might be indicative of a relative difficulty in suppressing an attractive but incorrect "OLD" response to F+/P- foils.

Overall, the analyses indicate that when surface features were absent, participants quickly rejected foils as "New," whereas when both surface features and principled properties were present participants accepted these items as "Old," although their responses were slower. However, when surface features were presents and principled properties absent (F+/P-), participants accurately rejected these items as "New," but their responses were significantly delayed as compared to all other items. This delay may be indicative of participants suppressing an attractive response of "Old" to F+P- items. In other words, for this foil there might be a competition between the two responses ("Old" vs. "New"), pointing to a relative difficulty for participants to suppress the salient surface feature and reject the foil. The fact that responses to F- foils were significantly faster than responses to F+ foils (1337 ms vs. 2521 ms) suggests that surface features and deep principled properties are accessed in a serial manner, with surface features being accessed first.

Experiment 3 establishes that novices encoded both surface and relational properties of the problems. The results also indicate that in the course of recognition, surface and deep features are accessed in a serial manner. However, the manner in which these features are encoded remains unclear. Specifically, there are two ways in which novices may encode these features of problems. First, it is possible that participants encoded both types of features in parallel. Second, it could be that participants encode them in a serial manner, attending to one of the features first (either surface or principled), and then to the other type of feature. The asymmetry in latencies to F- foils and P- foils (the former were rejected faster than the latter) indicates that the access is serial. The asymmetry also suggests that encoding is likely to be serial with surface features being encoded prior to principled properties. However, the asymmetry provides only suggestive evidence. To determine more conclusively which of these options actually takes place, we designed Experiment 4. Experiment 4 used the same recognition procedure as Experiment 3, with the exception that the length of presentation of stimuli in the study phase was reduced from 10 seconds to 1 second. If both types of features are encoded in parallel, then the patterns of accuracy among all foils should be identical to those in Experiment 3. If features are encoded in a serial manner, however, these patterns should be different. Specifically, if participants encode features first, then (when encoding time is limited) they should respond "NEW" to the foils in which surface features are absent, while responding "OLD" when surface features are present. Therefore, the proportion of "OLD" responses for the foil in which surface features are present but the principles are absent (F+/P-) should increase comparatively to the level of "OLD" responses observed in Experiment 3. If participants encode principled properties first, however, then the proportion of "OLD" responses for the foil in which surface features are absent but a principle is present (F-/P+) should increase comparatively to the proportions observed in Experiment 3.

EXPERIMENT 4

The goal of Experiment 4, which uses the same recognition procedure, as Experiment 3, is to determine whether participants encode features in a parallel or serial manner, and if the latter, then to determine the order in which features are encoded.

Method

Participants

Twenty-one undergraduates in an introductory psychology course at a large Midwestern university participated for course credit. Participants had an average age of 19.2 years (SD = 1.3 years), with 14 women and 7 men.

Materials and Procedure

The materials and procedures for this experiment were identical to those of Experiment 3, with one critical exception. In the current experiment, each of the thirty equations in the study phase was presented on the screen for one second, as opposed to ten seconds as in the previous experiment.

Results and Discussion

As for Experiment 3, for the current experiment we will first examine the number of OLD responses given for each of the five foils (i.e., correct acceptance of "Old" targets and incorrect acceptance of all other foils). After that we will specifically compare participants responses (in terms of accuracy and latency) to "Old" targets vs. F+/P- foils.

The percentages of OLD responses across foil types are presented in the left-hand side panel of Figure 5. These results indicate that the proportion of OLD responses for F-/P- (M = 0.08, SD = 0.12) and F-/P+ (M = 0.05, SD = 0.09) foils were significantly lower than chance (both ts(20) < -15.5, ps < .001), that the proportion of OLD responses for F+/P+ foils (M = 0.77, SD = 0.19) and Old Targets (M = 0.82, SD = 0.13) were significantly greater than chance (both ts(20) > 6.4, ps < . 001), and that the proportion of OLD responses for F+/P- foils (M = 0.56, SD = 0.28) was not significantly different from chance, t(20) < 1. These findings point to a number of important phenomena. First, participants were highly accurate at judging foils as NEW when surface features were absent (F-/P+ and F-/P- foils), indicating that even with a one-second exposure time, they were able to encode surface features. Second, when surface features were present but principles were absent (F+/P- foils), participants' judgments were at chance-levels, equally likely to consider the equation as OLD or NEW. Recall that in Experiment 3, these foils elicited approximately 80% of correct "NEW" responses. Finally, as in Experiment 3, because F+/P+ foils were categorically indistinguishable from Old Targets (since both surface features and principled properties present in Old Targets were also present in F+/P+ foils), participants most often considered these equations as OLD. Taken together, these results indicate that even with a one-second presentation of stimuli in the study phase, participants were able to encode the surface features, but were less likely to encode the principle properties.

A one-way repeated measures ANOVA points to significant differences in accuracy among foils ($\underline{F}$ (4, 80) = 70.41, $\underline{MSE}$ = .031, $\underline{p}$ < .001. Post-hoc comparisons using Bonferroni adjustments indicate that F-/P- and F-/P+ (correct answer "New") were most accurate (all $\underline{t}$s(20) > 3.2, all $\underline{p}$s < .05), whereas F+/P+ foils were least accurate (all $\underline{t}$s (20) < -3.1, all $\underline{p}$s < .05). In contrast to findings from Experiment 3, however, accuracy for F+/P- foils was significantly below that for Old Targets, $\underline{t}$ (20) = 5.6, $\underline{p}$ < .001, which again points to the fact that with the decreased encoding time, participants were less likely to encode principled properties and use these properties in their recognition judgments. These findings thus indicate that participants encode surface and principled features in a serial manner, with encoding for surface features occurring prior to that of principled properties.

The right-hand side of Figure 5 presents latencies to correct responses across F-/P+ ($\underline{M}$ = 1407 ms, $\underline{SD}$ = 1143 ms), F-/P- ($\underline{M}$ = 1328 ms, $\underline{SD}$ = 868 ms), F+/P- ($\underline{M}$ = 2400 ms, $\underline{SD}$ = 1616 ms), and OLD foils ($\underline{M}$ = 1634 ms, $\underline{SD}$ = 971 ms), and all responses for F+/P+ foil ($\underline{M}$ = 1930 ms, $\underline{SD}$ = 1076 ms). A one-way repeated measures ANOVA indicates significant differences among the foils, $\underline{F}$ (4, 392) = 60.3, $\underline{p}$ < .001. Post-hoc comparisons, using Bonferroni adjustments, were used to examine differences in latencies among foils. The patterns of differences among foils in the current experiment are qualitatively identical to those in Experiment 3. Again, latencies for F+/P- foils were significantly higher than those for Old Targets, $\underline{t}$(102) = 6.9, $\underline{p}$ < .001 and for F+/P+ foils, $\underline{t}$(106) = 6.6, $\underline{p}$ < .001, with no significant difference between latencies for Old targets and F+/P+ foils, $\underline{t}$(197) < 1. Also, the latencies of F-/P- and F-/P+ foils were significantly lower than those of the other three foils, all $\underline{t}$s > 7.1, all $\underline{p}$s< .001.

The results of Experiment 4 indicate that feature encoding is a serial process with surface features being encoded prior to deep relational properties. The results also corroborate findings of Experiment 3, indicating that accurate rejection of F+/P- items takes significantly more time than answer to any other item. Therefore, those participants who encoded both surface features and principled properties during the 1 second presentation time, exhibited the same pattern of competition between the "Old" and "New" responses as participants in Experiment 3. In both experiments, to produce the correct "New" response to F+/P- foils, participants needed to suppress the erroneous "Old" response. This competition between the two responses ("Old" vs. "New") points to a difficulty for participants to suppress the salient surface feature and reject the item. It seems likely that further reductions in exposure time should eliminate this competition because more participants would fail to encode principled properties.

GENERAL DISCUSSION

Several important findings stem from the four experiments presented in this paper. First, Experiment 1 indicates that experts consistently relied on principled properties, whereas novices relied mostly on surface features. Second, Experiment 2 indicates that novices exhibit a knowledge-representation dissociation, with most of the novices exhibiting knowledge of principled properties yet failing to represent these principles. Experiment 3 indicates that novices do encode principled properties, but there is a significant delay in responding whenever a surface feature is present and principled property is absent. Finally, Experiment 4 provides evidence that novices encode surface and principled features in a serial manner, with surface features encoded prior to principled properties. A hypothetical model of stimuli processing in novices is presented in Figure 6. The major components of the model are (1) that novices encode both surface features and deep principled properties and (2) whenever a surface feature is present, they have a tendency to accept the item as "Old," and they have to suppress this tendency when principled properties are absent. The model accounts for findings of Experiments 3 and 4. Indeed, when presented with F- foils, participants provided fast "NEW" responses. When presented with Old Targets and F+/P+ foils, participants provided "OLD" responses, which were slower than "NEW" responses to F- foils. Finally, when provided with F+/P- foils, participants provided "NEW" responses, which were slower than "NEW" responses to F- foils and than "OLD" responses to F+/P+ foils and Old Targets.

Of course, these findings raise an interesting question of whether or not experts would also exhibit such competition between "Old" and "New" answers in the case of F+/P- items. While this question was not within the scope of the present article, it will be specifically examined in future studies. Furthermore, an answer to this question (both a "yes" answer, that experts do exhibit the competition, or a "no" answer, that experts do not exhibit such a competition) would further our understanding of differences or similarities in problem encoding and problem representation by expert and novices.

These findings provide important information as to why experts and novices often approach problems in a different manner. As demonstrated in Experiment 1, expert-novice differences persist even when tasks are very simple, when principles are well familiar to novices, and when novices do not differ from experts in age or in overall ability. Therefore, we can eliminate possibilities that link the differences between experts and novices to overall ability or age differences. Results of Experiment 2 indicate that novices had knowledge of principled properties. Recall that when the impact of surface features was eliminated via fixing these features across the compared problems (in the "unmasked" phase of Experiment 2), most of the novices succeeded in construing principle-based representations. If they had not known the principled properties or had not been able to encode these features, the Unmasked condition should not differ dramatically from the both masked conditions. However, novices failed to construe a principle-based representation even when they knew principles in question.

Experiments 3 and 4 point to a putative mechanism underlying novices' failure to construe a principle-based representation. Recall that when surface features were present and principled properties were absent in Experiment 3 (i.e., the F+/P- foils), novices accurately judged these foils as "New." At the same time, the latencies of these responses were significantly higher than latencies to those foils where surface features were absent (i.e., the F-/P+ and F-/P- foils) or latencies of correct acceptance of Old Targets. Therefore, it takes time (and thus resources) to suppress a detected surface feature, and to answer "NEW." Furthermore, results of Experiments 3 and 4 indicate that surface features and principled properties are encoded and accessed serially, with the former being encoded and accessed prior to the latter. Thus it seems reasonable to infer that in Experiments 1 and 2 (where exposure to the problem was sufficient) novices initially encoded both surface and principled properties, but they failed to suppress more salient surface features in Masked phases of Experiments 1 and 2. In short, it seems that novices encode and discard principled properties, thus failing to focus on principles in Experiment 1 and the masked phase of Experiment 2. They fail to focus on principles because (a) they encode (and thus attend to) surface features prior to principled properties, and (b) it requires time and effort to suppress more salient surface features in favor of less salient principled properties.

Taken together, these results point to a plausible mechanism explaining previously found expert-novice differences (Chi, et al., 1981; Larkin, 1983; Larkin, et al., 1980; Simon & Simon, 1978). Furthermore, this mechanism is compatible with the literature on representation of surface and deep features in knowledge-lean domains. In particular, Gentner and her colleagues (Gentner & Medina, 1998; Kotovsky & Gentner, 1996; Gentner & Toupin, 1986) demonstrated that young children (considered by many as universal novices) typically prefer object matches to relational matches. For example, when presented with sets of objects similar to those in the first frame Figure 7, and asked what in the second set corresponds to the house in the first set, they typically answer that it is the house (Gentner & Medina, 1998). Gentner and her colleagues further demonstrated that it is possible to make even young children to notice relational similarity between the pictures when entities on the picture were sufficiently impoverished (second frame in Figure 7). In other words, when surface features are less attractive, participants are more likely to ignore them in favor of relational properties. These findings in conjunction with ours suggest that in grouping, categorization, and similarity judgment tasks, the more salient surface features are, the more likely it is that they win attentional competition over less salient relational properties. As a result, when presented with these tasks, participants are likely to discard relational properties (even when they properly encode these properties) in favor of more salient surface features.

Of course, it remains to be answered how people notice, extract, encode, and represent non-perceptual relations, those relations that can be used to solve a problem, and other abstract relational principles, such as the monotonic increase presented in Figure 1. One possibility is that relational words (e.g., "daughter, mother, grandmother" for the monotonic increase or "Boyle's law" for the invariance of volume-pressure product of a gas) facilitate such encoding and representation. Indeed, the ease with which even very young children notice relational similarity when relational language is provided (Lowenstein & Gentner, 1998; Kotovsky & Gentner, 1996), and physics experts notice relational similarity using conceptual entities of "force" and "energy" (Larkin, 1983) suggest that deep relations could be represented as concepts or schemas.

Therefore, the ability to represent deep relational properties may vary as a function of salience of competing surface features. These results seem to have important instructional implications. First, surface features (particularly those that are superfluous) may have larger negative impact on learning and performance than it was previously believed. In particular, superfluous surface features do not appear to be merely irrelevant; they tend to distract attention from more important principled properties. Because in novices the suppression of surface features in favor of deep relational properties requires effort, novices should benefit when superfluous surface features are de-emphasized. This decrease in attentional weights of surface features may lead to a simultaneous increase in attentional weights of deep principled properties.

Results, however, should not be overgeneralized. While they tell a part of the story, they do not tell the whole story. In particular, it remains unknown whether or not simplified tasks and well-familiar deep principles are fully representative of more complex tasks and less-familiar deep principles. If the task and principles are not fully representative of their more complex counterparts, there is a possibility that alternative factors, such as knowledge or availability due to practice, could play a role when the complexity increases. The major outcome of this paper is that expert-novice differences persist even under most stripped down and simplified task conditions and that under these conditions novices fail to represents deep relational principles. This failure appears to be due to greater attentional demands required to encode these features, and effort required to suppress more salient surface features in favor of less salient principled properties. Whether or not this mechanism accounts for novice performance under more complex task conditions requires further research.

Present results allow us to conclude that under these task conditions, the observed problem representations in novices do not stem from inadequate knowledge or low general ability, but they rather stems from attentional competition between more salient surface features and less salient deep relational principles.

References

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. Journal of Experimental Psychology: Learning, Memory, & Cognition, 10, 483-495.

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.

Bassok, M. (1996). Using content to interpret structure: Effects on analogical transfer. Current Directions in Psychological Science, 5, 54-58.

Bassok, M. (1997). Two types of reliance on correlations between content and structure in reasoning about word problems. In L. D. English (Ed.), Mathematical reasoning: Analogies, metaphors, and images, (pp. 221-246). Mahwah: NJ: Lawrence Erlbaum Associates.

Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. Journal of Experimental Psychology: Learning, Memory, & Cognition, 22, 792-810.

Brown, A. L. (1989). Analogical learning and transfer: What develops? In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning, (pp. 369-412). New York: Cambridge University Press.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. Cognition. 31(1), 61-83.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. Cognitive Psychology, 17, 391-416.

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

Custers, E., Boshuizen, H., & Schmidt, H. G. (1998). The role of illness scripts in the development of medical diagnostic expertise: Results from an interview study. Cognition & Instruction, 16(4), 367-398.

Evans, J. St. B. T., Newstead, S. E.; Byrne, R. M. J. (1993). Human reasoning: The psychology of deduction. Hove, England: Lawrence Erlbaum Associates.

Everyday Mathematics: Teacher's Reference Manual (1998). Chicago, IL: Everyday Learning.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. Cognition, 23, 183-209.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning, (pp. 199-241). New York: Cambridge University Press.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. Cognition, 65, 263-297.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. Cognitive Science, 10, 277-300.

Hinsley, D. A., Hayes, J. R. & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In P. A. Carpenter & M. A. Just (Eds.), Cognitive processes in comprehension (pp. 89-106). Hillsdale, NJ: Erlbaum.

Holyoak, K. J.; Koh, K. (1987). Surface and structural similarity in analogical transfer. Memory & Cognition, 15, 332-340.

Horn, J. & Noll, J. (1994). A system for understanding cognitive capabilities: A theory and evidence on which it is based. In D. Detterman (Ed.), Current topics in human intelligence: Volume 4. Theories of intelligence (pp. 151-203). Norwood, NJ: Ablex.

Johnson-Laird, P., & Byrne, R. (1991). Deduction. Hove, UK: Lawrence Erlbaum.

Kaiser, M. K., McCloskey, M., & Proffitt, D. R. (1986). Development of intuitive theories of motion: Curvilinear motion in the absence of external forces. Developmental Psychology, 22, 67-71.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.

Kyllonen, P. C. (1994). CAM: A theoretical framework for cognitive abilties measurement. In D. Detterman (Ed.), Current topics in human intelligence: Volume 4. Theories of intelligence (pp. 307-359). Norwood, NJ: Ablex.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. Cognitive Science, 4, 317-345.

Larkin, J. (1983). The role of problem representation in physics. In D. Gentner & A. Stevens (Eds.), Mental models, (pp. 75-98). Hillsdale, NJ: Erlbaum.

Lowenstein, J., & Gentner, D. (1998, August). Relational language facilitates analogy on children. Paper presented at the 20th Meeting of the Cognitive Science Society. Madison, Wisconsin.

McClosskey, M. (1983). Naïve theories of motion. In D. Gentner & A. Stevens (Eds.), Mental models, (pp. 289-324). Hillsdale, NJ: Erlbaum.

Morris, A. K., & Sloutsky, V. M. (1998). Understanding of logical necessity: Developmental antecedents and cognitive consequences. Child Development, 69, 721-741.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. Science, 238, 625-631.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. Journal of Experimental Psychology: Learning, Memory, & Cognition, 14, 510-520.

Reed, S. K., Ackinclose, C. C., & Voss, A. A. (1990). Selecting analogous problems: Similarity versus inclusiveness. Memory & Cognition, 18, 83-98.

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. Psychological Bulletin, 115, 381-400.

Ross, B. H., & Kilbane, M. C. (1997). Effects of principle explanation and superficial similarity on analogical mapping in problem solving. Journal of Experimental Psychology: Learning, Memory, & Cognition, 23, 427-440.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. Journal of Experimental Psychology: Learning, Memory, & Cognition, 8, 484-494.

Siegler, R. S. (1978). The origins of scientific thinking. In R. S. Siegler, (Ed), Children's thinking: What develops? (pp. 109-149). Hillsdale, NJ: Lawrence Erlbaum Associates.

Silver, E. A. (1981). Recall of mathematical problem information: Solving related problems. Journal of Research in Mathematics Education, 24, 117-135.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler, (Ed), Children's thinking: What develops? (pp. 325-348). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. New York: Cambridge University Press.

Author Note

This research was supported in part by a grant from James S. McDonnell Foundation. We would like to thank Martha W. Alibali, Brad Morris and Aaron Rader for helpful comments. Johnny Chuah's assistance in running participants is greatly appreciated.

Correspondence concerning the manuscript should be addressed to Vladimir M. Sloutsky (sloutsky.1@osu.edu) or Aaron S. Yarlas (yarlas.1@osu.edu), Center for Cognitive Science, 21 Page Hall, 1810 College Avenue, The Ohio State University, Columbus, OH 43210.

Table 1. Problem templates used for the nine feature comparisons in Experiment 1.

	Template		
Test-Target Comparison	Target	Test 1	Test 2
Number vs. Associativity	a + b + c = [a+x - x] + b + c	a + b + c = d + e + f (N)	g + h + i = [g+y - y] + h + i (A)
Sign vs. Associativity	a + b = [a+x - x] + b	d + e = f + g + h (S)	i – j = [i+y - y] – j (A)
NOE vs. Associativity	a + b = [a+x - x] + b	d + e = f + g + h (NOE)	i + j + k = [i+y - y] + j + k (A)
Number vs. Commutativity	a + b + c = c + a + b	a + b + d = c + b + e (N)	f + g + h = h + f + g (C)
Sign vs. Commutativity	a + b = b + a	c + d = e + f (S)	g * h = h * g (C)
NOE vs. Commutativity	a + b + c = c + a + b	d + e + f = g + h + i (NOE)	j + k = k + j (C)
Number vs. Sign	a + b = c + d	a – b = c – e (N)	f + g = h + i (S)
Number vs. NOE	a + b = c + d	a + b + e = c + d + e (N)	f + g = h + i (NOE)
Sign vs. NOE	a + b = c + d	e + f + g = h + i +j (S)	k – l = m – n (NOE)

Note: A = associativity, C = commutativity, N = numbers, NOE = number of elements, S = sign. All lowercase letters indicate numbers.

Table 2. Percentage of participants in each sample who made explanation-consistent principled choices never, inconsistently, or consistently in Experiment 1.

	Use of explanation-consistent principled choices		
Sample	Never	Inconsistently	Consistently
Undergraduates	52.00	40.00	8.00
History graduate students	37.50	50.00	12.50
Mathematics graduate students	10.00	15.00	75.00

Note. Never = made 0 out of 18 explanation-consistent principled choices, Inconsistently = made 1 to 13 out of 18 explanation-consistent principled choices, Consistently = made 14 to 18 explanation-consistent principled choices

Figure Captions

Figure 1. Example of a matching task, where Test 2 represents an object match with the Target, whereas Test 1 represents relational a relational match.

Figure 2. Percentage of explanation-consistent choices for each principle for Unmasked and Masked trials in Experiment 2, and for Masked trials in Experiment 1. Error bars represent Standard Errors of the Mean.

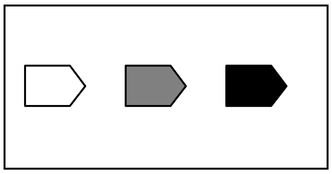
Figure 3. Patterns of answer when (1) only features are encoded and (2) both features and principles are encoded.

Figure 4. Percentages of "OLD" responses and response times (in milliseconds) across foil types in the recognition phase of Experiment 3. Error bars represent Standard Errors of the Mean. F+/P- = Feature +/Principle -, F+/P+ = Feature +/Principle +, F-/P- = Feature -/Principle -, F-/P+ = Feature - /Principle +, and OLD = Old Targets.

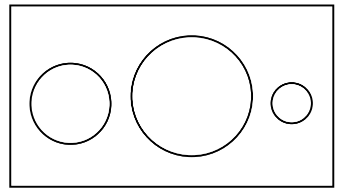
Figure 5. Percentages of "OLD" responses and response times (in milliseconds) across foil types in the recognition phase of Experiment 4. Error bars represent Standard Errors of the Mean. F+/P- = Feature +/Principle -, F+/P+ = Feature +/Principle +, F-/P- = Feature -/Principle -, F-/P+ = Feature - /Principle +, and OLD = Old Targets.

Figure 6. Proposed flow-model of processing mechanisms underlying recognition based on data from Experiments 3 and 4.

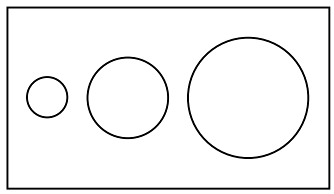
Figure 7. Informationally-rich and informationally-impoverished objects used by Gentner & Medina (1998).



Test 1

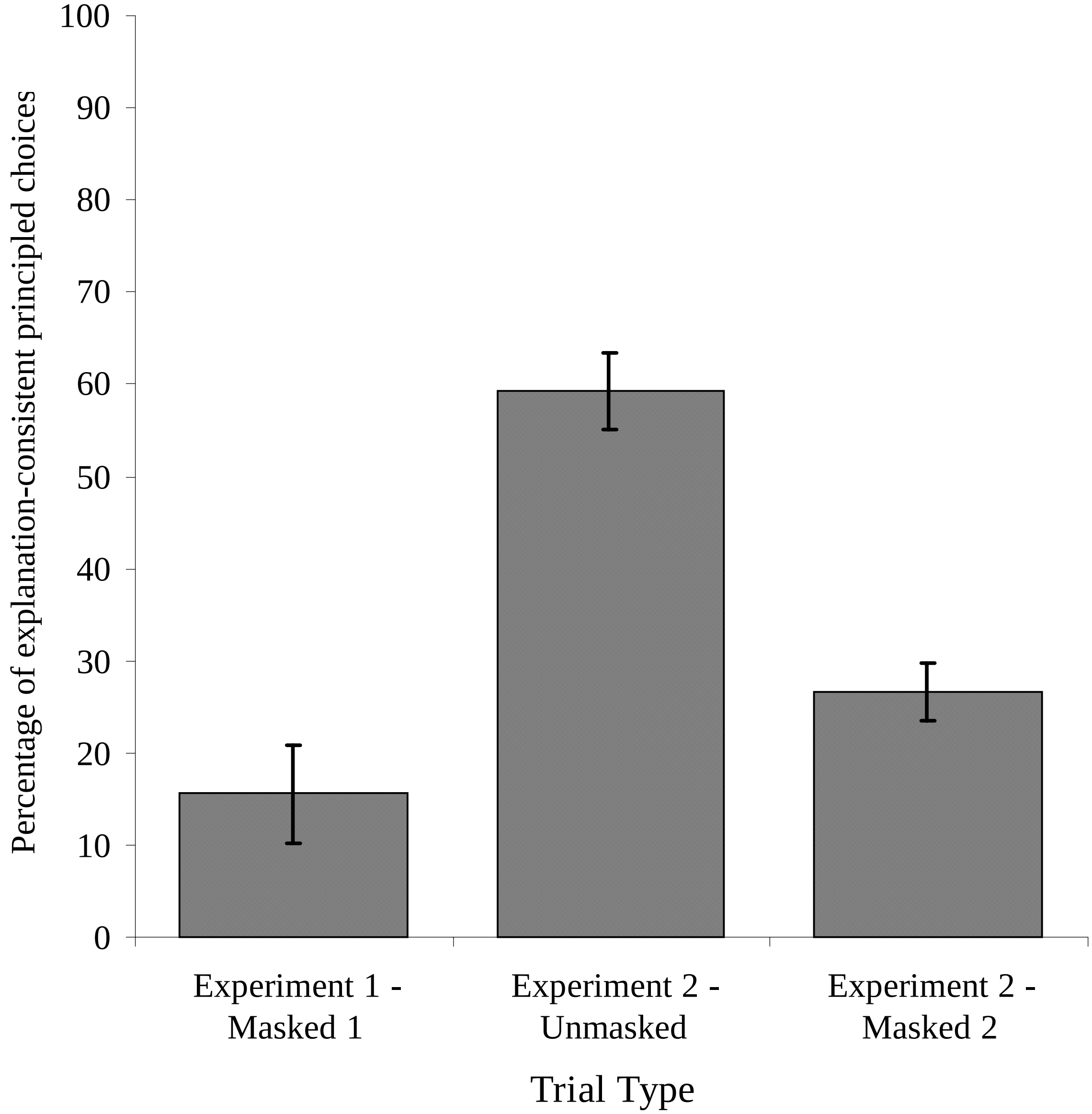


Test 2



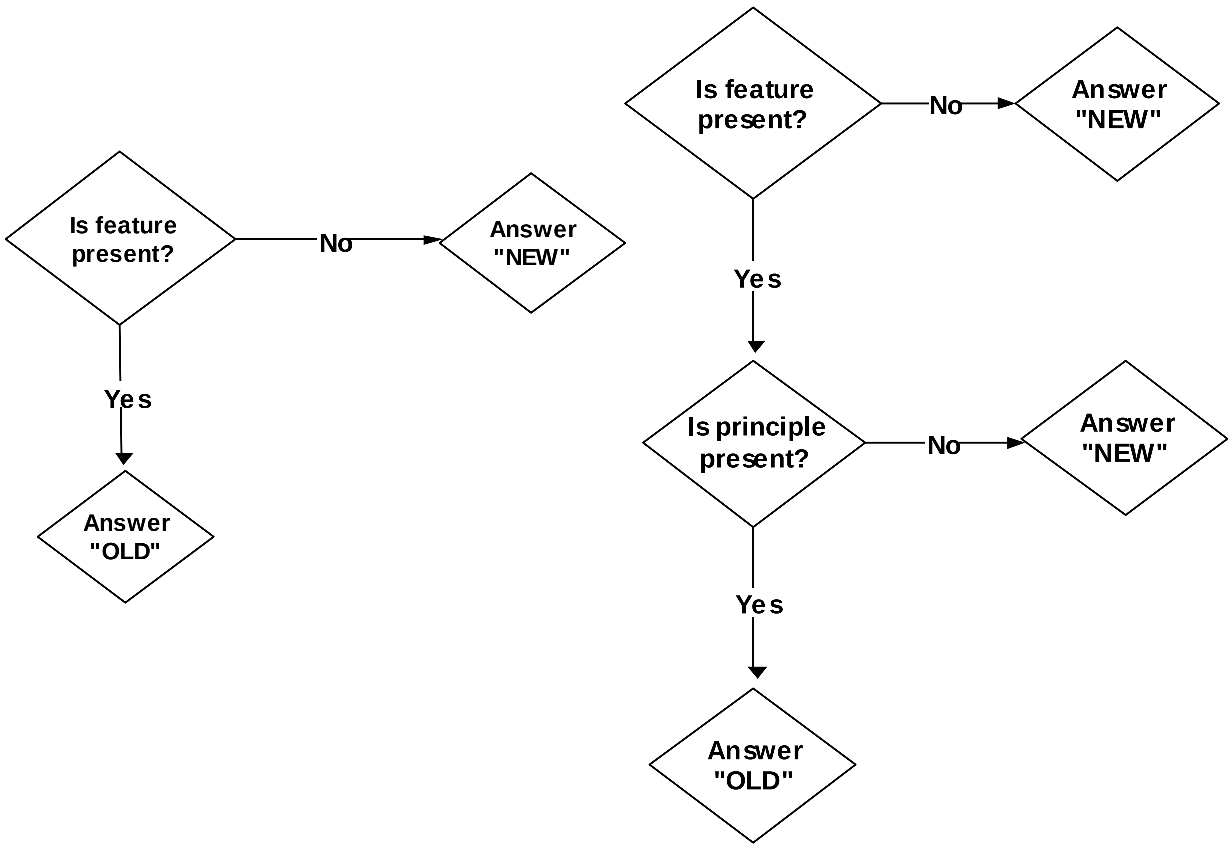
Target

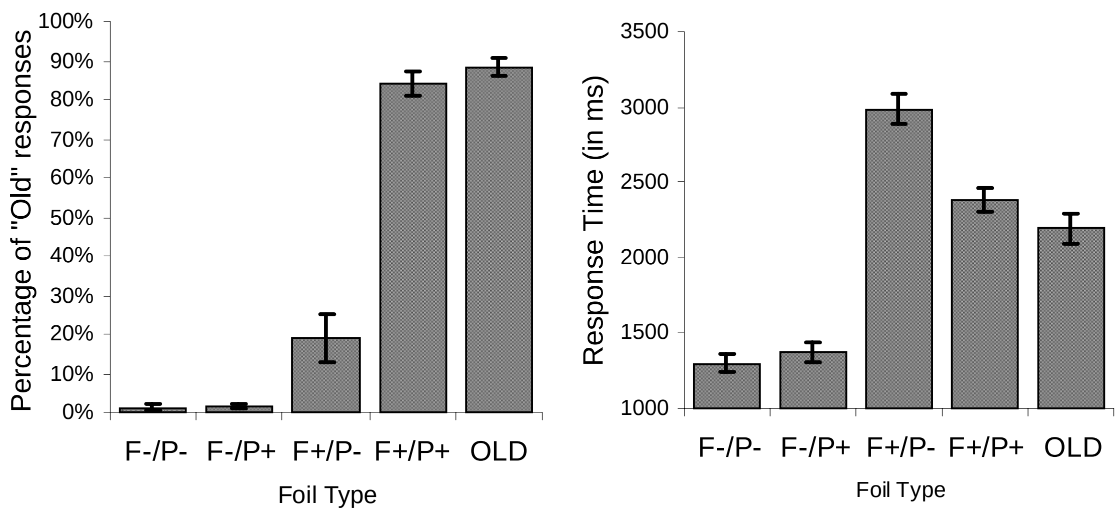
Which of the Test stimuli is more similar to the Target?


Percentage of explanation-consistent principled choices
100
90
80
70
60
50
40
30
20
10
0
Experiment 1 - Masked 1
Experiment 2 - Unmasked
Experiment 2 - Masked 2
Trial Type

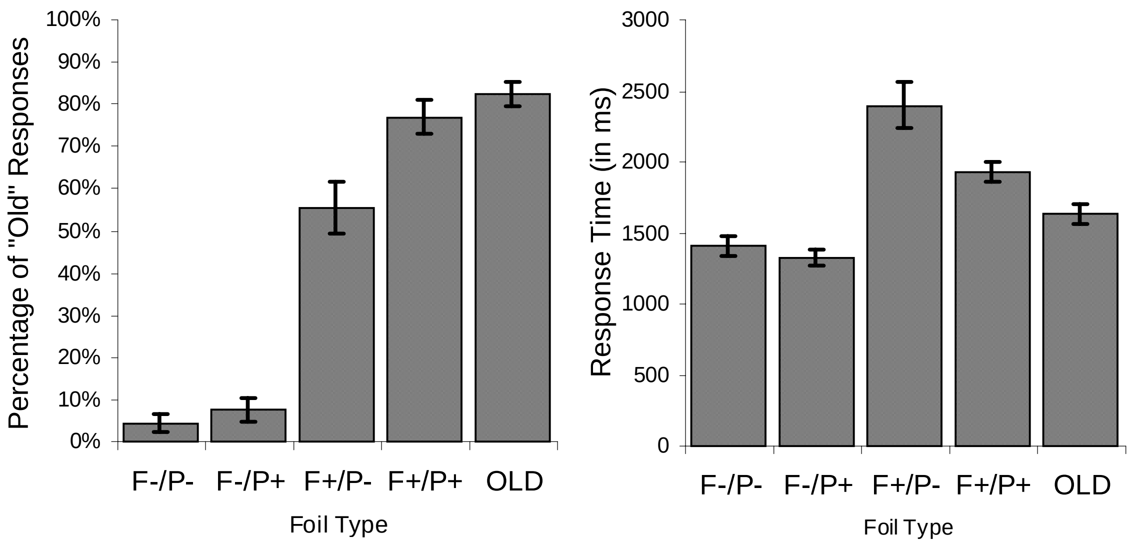
Possibility 1
Only features are encoded

Possibility 2
Both features and principles are encoded

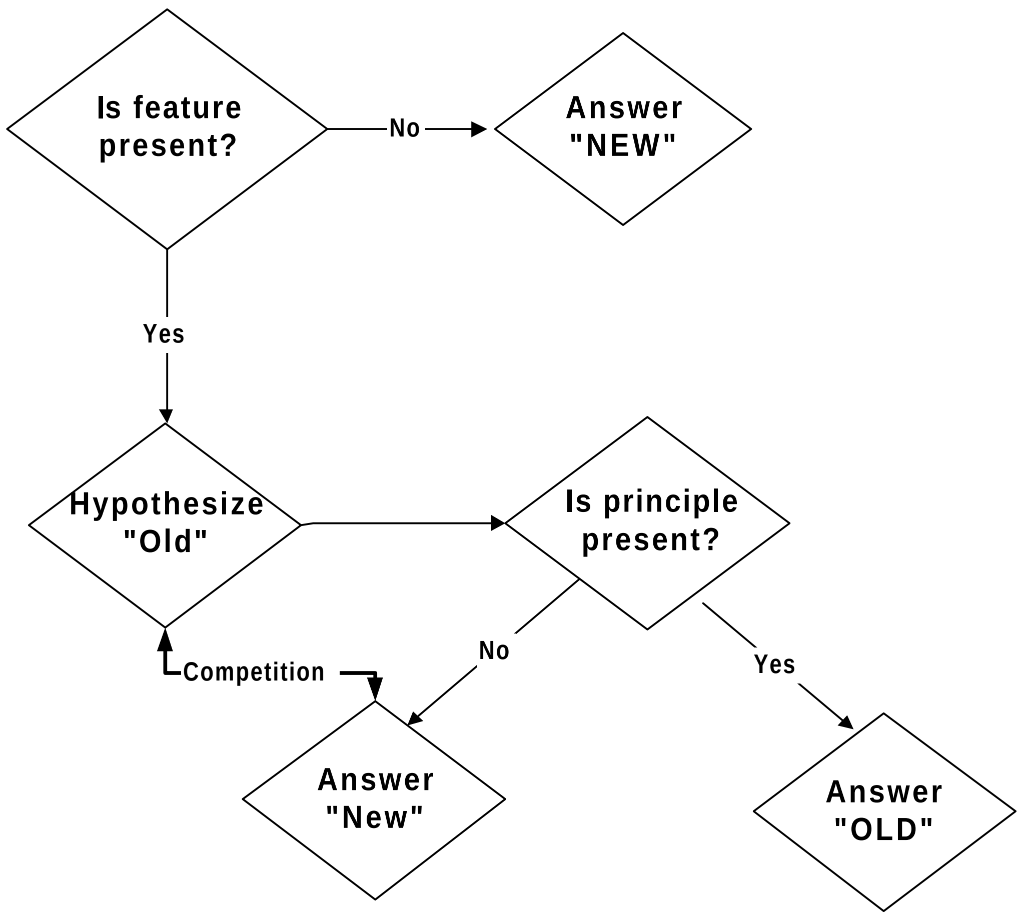




Percentage of "Old" responses
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
F-/P-
F-/P+
F+/P-
F+/P+
OLD
Foil Type
Response Time (in ms)
3500
3000
2500
2000
1500
1000
F-/P-
F-/P+
F+/P-
F+/P+
OLD
Foil Type



100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of "Old" Responses
F-/P-
F-/P+
F+/P-
F+/P+
OLD
Foil Type
3000
2500
2000
1500
1000
500
0
Response Time (in ms)
F-/P-
F-/P+
F+/P-
F+/P+
OLD
Foil Type


Is feature present?
No
Answer "NEW"
Yes
Hypothesize "Old"
Is principle present?
No
Yes
Competition
Answer "New"
Answer "OLD"