

Investigating Test Items Designed to Measure Higher-Order Reasoning using Think-Aloud Methods: Implications for Construct Validity and Alignment

Jacqueline P. Leighton

Rebecca J. Gokiert

Centre for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta

Paper presented at the Annual Meeting of the American Educational Research Association (AERA), Montreal, Quebec, Canada (April, 2005).

Abstract

Given that categorical and conditional tasks are used in varying forms on many standardized tests to measure higher-level reasoning (Powers & Dwyer, 2003), we speculated on whether these reasoning tasks might contain sources of ambiguity, leading to misalignment. Sixteen undergraduate students (8 females and 8 males; \bar{x} age = 20 years and 3 months; SD = 1 year 9 months) enrolled in an introductory symbolic logic course at a large research-intensive university were recruited to participate in the present study. Each of the 16 students was interviewed twice—once within the first week of the course and then again within the last week of the course. The purpose of the interviews was to compare their cognitive processing and performance on categorical and conditional syllogisms at two different stages of knowledge acquisition. Concurrent think-aloud and retrospective think-aloud interviews were used. Results indicate that test items are aligned with students' interpretations of difficulty. Retrospective reports supported by concurrent reports revealed a source of misalignment between students and tasks. The source of this misalignment originated from students who were not devoting enough time to fully comprehending important conceptual features of the tasks they were solving. This lack of time resulted in improper application of strategies to solve tasks successfully.

Investigating Test Items Designed to Measure Higher-Order Reasoning using Think-Aloud Methods: Implications for Construct Validity and Alignment

The alignment between test items and how students understand or interpret the test items has become increasingly important in establishing the construct validity of a test (Aikenhead, 1988; Aikenhead & Ryan, 1992; Ercikan et al., 2004; Ferrara et al., 2003, 2004; Haladyna & Downing, 2004; Standards, 1999; National Research Council, 2001). As educational tests are developed to measure higher-level cognitive processes such as analytical or causal reasoning and multi-step problem solving (National Research Council, 2001), it is necessary to investigate the cognitive processes students use when responding to these test items (Leighton, 2004). Therefore, the purpose of the present research was to (a) investigate the alignment of categorical and conditional syllogisms for measuring higher-level reasoning in students and (b) demonstrate the feasibility of using both concurrent and retrospective think-aloud methods for collecting evidence of students' cognitive processing on reasoning test items (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995; Taylor & Dionne, 2000).

Reasoning Tasks: Are They Really Measuring Reasoning?

In an effort to develop test items that measure more sophisticated forms of cognitive processing such as reasoning, test developers on occasion have borrowed cognitive tasks from the psychological literature (Irvine & Kyllonen, 2002). For example, object assembly tasks and matrix completion problems have been borrowed to develop spatial reasoning and abstract reasoning test items (e.g., Embretson, 2002; Embretson & Gorin, 2001). Furthermore, various forms of categorical and conditional syllogisms have been borrowed to measure analytical reasoning in tests such as the SAT I: Reasoning Test, GRE, GMAT, and LSAT (Powers & Dwyer, 2003).

As shown in Figure 1, categorical and conditional syllogisms usually involve two or more statements (premises) that a student must consider together in order to generate a conclusion. Within cognitive psychology, there has been a long standing assumption that categorical and conditional syllogisms elicit higher-order or logical reasoning (Irvine & Kyllonen, 2002; see also Stanovich, 1999). However, more recently, cognitive psychologists concerned about the validity of these tasks, have increasingly begun to question this assumption (Fiddick, Cosmides, & Tooby, 2000; Girotto, 2004; Leighton & Sternberg, 2003; Roberts & Newton, 2005). Educational measurement researchers and practitioners must exercise caution when borrowing cognitive tasks from the cognitive psychological domain. Cognitive tasks borrowed from the psychological research domain must be subjected to the same kinds of construct validity investigations as other educational test items. Construct validity investigations should especially focus on the alignment between test items and students.

Ferrara and his colleagues (2004, p. 1) define the alignment between test items and students as “the degree of correspondence between content area knowledge, content area skills, broader cognitive processes, and response strategies (i.e., Knowledge, Skills, Processes, and Strategies: KSPS) that (a) test developers intend to assess, and (b) examinees bring to bear when they respond to test items.” When the knowledge and skills test developers intend to assess correspond with the observed knowledge and skills students apply to correctly answer test items, then an item is *aligned* with student interpretations. Conversely, when the knowledge and skills test developers intend to assess do not correspond with the observed knowledge and skills students apply to correctly answer test items, then an item is *misaligned* with student interpretations (see also Ferrara et al., 2003).

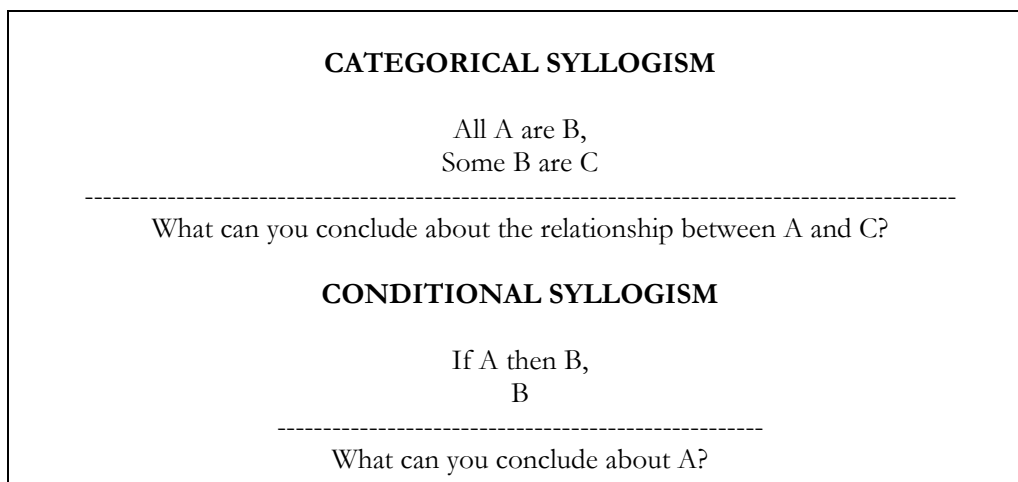


Figure 1. Examples of categorical and conditional syllogisms

There is research evidence to suggest that the alignment between item objectives and student interpretations is often questionable (Munby, 1982). A survey of the most recent developments in the psychology of reasoning indicates that the alignment between how individuals mentally represent (or understand) cognitive tasks and how psychologists *expect* individuals to understand the tasks is surprisingly different (Fiddick et al., 2000; Girotto, 2004; Leighton & Sternberg, 2003). This discrepancy can undermine the validity of the inferences made about individual test performance (see Ercikan et al., 2004; Ferrara et al., 2003, 2004; Haladyna & Downing, 2004; Leighton, 2004; Leighton & Sternberg, 2003; Roberts & Newton, 2005).

In the field of science education, studies indicate that students and test developers do not necessarily perceive the same meaning in test items (Aikenhead, 1988; Aikenhead & Ryan, 1992; Ferrara et al., 2003, 2004). For example, in their comparison of Likert-type and written paragraph responses from the International Association for the Evaluation of Educational Assessment (IEA) science study, Aikenhead, Fleming, and Ryan (1987) found that the way in which students view science was not accurately measured (see also Brunkhorst, 1987). Yaroch (1986) discovered a consistent misrepresentation of students' biological ideas when students were evaluated by a standardized multiple-choice test. In addition, Ferrara et al. (2004) recently found that in a state-wide standardized science assessment, two out of 20 publicly released items were completely misaligned and another 11 items were partially misaligned. Ferrara et al.'s conclusions were that seemingly simple words and phrase choices in test items could (a) disrupt how students understand the nature of the task, (b) derail students' cognitive processing, and (c) undermine the construct validity of individual items and the test as a whole.

In a recent study (Leighton, under review) investigating the influence of symbolic logic instruction on the assessment of higher-level reasoning, two groups of students (control and comparison groups) were assessed on 16 categorical and 16 conditional reasoning tasks of increasing difficulty using a paper-and-pencil format. The control group (100 students) received no instruction in logic, while the comparison group (approximately 120 students) received 12-weeks of symbolic logic instruction by a tenured professor. One of the results from this study indicated that training students in logic over 12-weeks increased students' performance on difficult reasoning tasks only by a modest amount. Although trained students improved their performance over the course of 12 weeks, and performed better than untrained students, especially on selected-response tasks, trained students still performed poorly on difficult reasoning tasks (scores generally fell below 50%). Given that the content of instruction *matched* the content of the assessment tasks, transfer of training was

expected to be high (see Sternberg & Ben-Zeev, 2001) and it was therefore puzzling to not observe a greater improvement in performance. Although other researchers had obtained similar findings in past studies and concluded that certain forms of reasoning are not easily taught (Cheng, Holyoak, Nisbett, & Oliver, 1986; Lehman, Lempert, & Nisbett, 1988; Morris & Nisbett, 1993), further investigation seemed necessary of the strategies students were selecting and applying as they solved the categorical and conditional tasks. Given that categorical and conditional tasks are used in varying forms on many standardized tests to measure higher-level reasoning (Powers & Dwyer, 2003), we speculated on whether these reasoning tasks might contain sources of ambiguity, leading to misalignment.

Method and Procedure

Participants. Sixteen undergraduate students (8 females and 8 males; \bar{x} age = 20 years and 3 months; SD= 1 year 9 months) enrolled in an introductory symbolic logic course at a large research-intensive university were recruited to participate in the present study. The instructor of the course was an experienced tenured professor in Symbolic Logic who used the text book by Virginia Klenk, (2002), *Understanding symbolic logic 4th edition*. NJ: Prentice Hall.

Think aloud interviews. Each of the 16 students was interviewed twice—once within the first week of the course and then again within the last week of the course. The purpose of testing students twice (using identical interview procedures) was to compare their cognitive processing and performance on categorical and conditional syllogisms at two different stages of knowledge acquisition. At the beginning of the course, students lacked formal training in symbolic logic and this lack of experience was expected to reveal itself in the verbal reports. In contrast, at the end of the course, students were expected to have learned some of the basic vocabulary and strategies of symbolic logic and were anticipated to reveal this emerging expertise in their verbal reports. Participating students represented a range of educational backgrounds and had no previous background in symbolic logic.

Instructions. The authors conducted all interviews. After familiarizing students to the nature of the tasks and think-aloud procedures (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995), students were given four categorical syllogisms to solve and four conditional syllogisms to solve of varying difficulty in either a constructed-response (CR) or selected-response (SR) format. Students were not told which tasks were easy or difficult. Tasks were also counter-balanced. The task instructions shown below are considered standard instructions for constructed-response categorical syllogisms and were used in the present study (adapted from Johnson-Laird & Bara, 1984):

In the following pages, you will see 4 pairs of statements about different people or groups of people, whom you should imagine as assembled in a room. After reading each pair of statements, please write down what, if anything, follows necessarily from these premises about the occupants of the room. If you consider that there is no conclusion that follows necessarily from the premises, please write down "Nothing." Your conclusions should be based solely on the information presented in the statements, and NOT on plausible suppositions or general knowledge. For example, please read the statements below:

1. *All artists are beekeepers*
All beekeepers are chefs
Then...All artists are chefs

From these statements, I can definitely conclude that All artists are chefs so I write this down below the statements.

2. *Some acrobats are bakers*
Some bakers are canoeists
Then...Nothing

From these statements, I can NOT conclude anything definite about acrobats and canoeists so I write down "nothing" below the statements. If you have any questions, please ask the experimenter now. If you don't have any questions, please begin task 1 on the next page. Please, do NOT skip any questions.

The task instructions for the selected-response categorical condition were identical to the instructions above except that five alternatives followed examples 1 and 2 (for example 1: All artists are chefs, some artists are chefs, some artists are not chefs, no artists are chefs, and nothing). The task instructions for the conditional syllogisms patterned after Johnson-Laird and Bara's (1984) instructions were as follows:

In the following pages, you will see 4 pairs of statements about people located in well-known places. After reading the pair of statements, please write down what, if anything, follows necessarily from these premises about the location of one of the persons. If you consider that there is no conclusion that follows necessarily from the premises, please write down "Nothing." Your conclusions should be based solely on the information presented in the statements, and NOT on plausible suppositions or general knowledge. For example, please read the statements below:

1. *If Hank is in Chicago, then Julia is in Baghdad*
Hank is in Chicago
Then... Julia is in Baghdad

From these statements, I can definitely conclude that Julia is in Baghdad so I write this down below the statements.

2. *If Linda is in Madrid, then Robert is NOT in Philadelphia*
Linda is NOT in Madrid
Then...Nothing

From these statements, I can NOT definitely conclude that Robert is NOT in Philadelphia so I write down "nothing" below the statements. If you have any questions, please ask the experimenter now. If you don't have any questions, please begin task 2 on the next page. Please, do NOT skip any questions.

Each set of categorical and conditional syllogisms contained problems at two levels of difficulty (low and high). In the concurrent portion of the interview, students were reminded to "keep talking" as they solved the tasks. After finishing with the concurrent portion of the interview, students were asked a series of retrospective questions aimed at confirming the information provided in the concurrent reports and identifying students' metacognitive knowledge (Taylor & Dionne, 2000). Metacognitive knowledge is knowledge about global problem-solving strategies that guide more specific, local strategies. Metacognitive knowledge has been found to be useful in revealing students'

general level of task understanding and factors underlying students' performance success or failure (Kuhn, 2001). The retrospective questions included the following:

- (a) For the first set of tasks, which problem or problems did you find most difficult and why?
- (b) What kind of strategies do you think you used as you solved the first set of tasks?
- (c) For the second set of tasks, which problem or problems did you find most difficult and why?
- (d) What kind of strategies do you think you used as you solved the second set of tasks?

Students' verbal reports were summarized using a technique developed for another, larger study in the domain of science reasoning (see Leighton & Gokiert, 2005). This technique involves identifying prominent themes in students' reports by identifying students' repetition of key item words, supplementary words or phrases, and strategies.

Results and Discussion

Students' Performance on Categorical and Conditional Syllogisms

Descriptive statistics for students' performance on categorical and conditional syllogisms are shown in Table 1. Please note that the mean scores within the table represent values out of 4.

Table 1
Mean Scores and Standard Deviations for Categorical and Conditional Syllogisms by Time in Course and Item Format

Time 1				Time 2			
Categorical		Conditional		Categorical		Conditional	
CR	SR	CR	SR	CR	SR	CR	SR
1.625	2.563	3.188	2.562	1.750	2.750	3.500	3.125
<i>[40.6%]</i>	<i>[64.1%]</i>	<i>[79.7%]</i>	<i>[64.1%]</i>	<i>[43.8%]</i>	<i>[68.8%]</i>	<i>[87.5%]</i>	<i>[78.1%]</i>
(0.518)	(0.980)	(0.843)	(1.294)	(0.707)	(1.035)	(0.535)	(0.991)

Note. CR = Constructed Response; SR = Selected Response. Mean percentage shown in italics in square brackets; Standard deviations are presented within parentheses

A 2 (task format) by 2 (gender) by 2 (syllogism type) by 2 (time) within-subject ANOVA with repeated measures on the last two factors was performed. Similar to the results found in a previous study (see Leighton, under review), there were three significant findings in the present study. First, there was a significant main effect of syllogism type ($MS_e=13.598$, $F(1, 12)=8.652$, $p=.012$), indicating that students performed significantly better on conditional syllogisms than on categorical syllogisms. This finding replicates a previous result using a larger sample size (Leighton, under review). Second, there was a significant interaction between syllogism type and item format ($MS_e=8.629$, $F(1, 12)=5.490$, $p=.037$), indicating that students performed significantly better on selected-response categorical syllogisms than on constructed-response categorical syllogisms but the reverse was true on conditional syllogisms; students performed better on constructed-response conditional syllogisms than on selected-response conditional syllogisms. Again, this result replicates

a previous finding using a larger sample size (Leighton, under review). Third, there was a significant main effect of time ($MS_e=1.410$, $F(1, 12)=6.333$, $p=.027$), indicating that students generally performed better on syllogisms after their 12-week course in logic. The results found with the sample of 16 students replicates the results found in another study of students' performance on logical reasoning tasks (see Leighton, under review).

Students' final reported grades in the symbolic logic course were positively but not significantly correlated with performance on the reasoning tasks. As shown in Table 1, students' overall performance on the syllogisms is low, especially performance on CR categorical syllogisms with students receiving less than 50% of items correct at both time 1 and time 2. As mentioned previously, we sought to investigate the strategies students were selecting and applying as they solved the categorical and conditional tasks to better understand their overall modest performance.

Students' Concurrent and Retrospective Verbal Reports

Concurrent reports at Times 1 and 2. Three themes were identified when reviewing the concurrent reports. First, students who completed the syllogisms in the selected-response format *made use of the options* (alternatives) to help them evaluate possible answers to the syllogisms. Six out of eight students who completed the tasks in a SR format made repeated reference to the options as they solved the categorical and conditional tasks during time 1. By time 2, the reference to options was reduced. Only one student explicitly mentioned using the options at time 2. That most students did not explicitly mention the options at time 2 does not unequivocally suggest that students did not actually use the options in their reasoning. However, the lack of mention at time 2 suggests that the options might have played a lesser role in their reasoning.

The verbal report from K.S. is illustrative of the way in which students methodically reviewed task options in their approach to both sets of syllogisms in the SR format at time 1:

K.S.: "Some librarians are travelers, all travelers are counselors." All librarians are counselors – not necessarily because some librarians are not travelers, and some librarians are counselors, that's true because some are travelers and all travelers are counselors given these premises. Some librarians are not counselors, that's true as well because some librarians aren't travelers, making them not counselors. And no librarians are counselors, given these premises, that's not true because they told us that all travelers are counselors.

A second theme in students' concurrent reports at times 1 and 2 was the *prevalence of a single goal*. Fourteen out of 16 students or approximately 88% of students attempted to generate a *link* or connection *between the individuals or sets of individuals* mentioned in the premises of the syllogisms as they reasoned through the syllogisms. This attempt at establishing a link is noteworthy because the ability to generate, use, and/or evaluate a link between premises is one of the skills being measured by categorical and conditional syllogisms. According to psychologists, the understanding of this link or implication between premises and conclusion is essential to generating sound conclusions to logical tasks and is a concept that is required in logical and higher-level reasoning (see Johnson-Laird & Bara, 1984; Powers & Dwyer, 2003; Stanovich, 1999).

According to students' concurrent reports, the cognitive effort expended by students to achieve this goal appeared to be greater for categorical syllogisms than conditional syllogisms (see also retrospective portion below). The effort expended appeared greater for categorical syllogisms because the connection or link between premises and conclusion must be generated for *sets of individuals* (as per the quantifiers all, some, some...not, and no). In contrast, the effort to establish a

link appeared to be lessened for conditional syllogisms because the link applies only to two individuals (the two individuals mentioned in the premises of the conditional syllogism). Moreover, the link is largely established in the conditional rule (one of the premises of the conditional), and is not muddled by the need to think of sets of individuals. As a result, the reasoning process applied to conditional tasks appeared more straightforward in the concurrent reports because students used the statements of the conditional *directly* to generate a response without reflecting on any sets. In fact, students simply repeated the conditional premises and from these statements generate a response. Consider A.L.'s report as he reasons through a conditional syllogism at time 1:

A.L.: “If Jane is in Paris, then Maria is in Rome. Jane is in Paris”....so then Maria is in Rome as per the first sentence. “If Meigal is in Venice, then Polly is in Texas. Polly is in Texas” it says nothing about Meigal, so we cannot conclude about Meigal. “If Jackie is in Toronto, then Tony is in Washington. Jackie is not in Toronto”. That does not follow that Tony is in Washington I don't think. Since it only states that Jackie is in Toronto and Tony is....so nothing. “If Josh is in Rio, then Margaret is in Lima. Margaret is not in Lima”. Hm....I know I'm....

Now consider the verbal report A.L. provides as he reasons through a categorical syllogism at time 1:

A.L.: “Some veterinarians are not runners, some housekeepers are runners”. Veterinarians and housekeepers and the reference to runners have nothing in common here if veterinarians are not runners and housekeepers are runners, it does not say anything between them, so nothing. “Some librarians are travelers, all travelers are counselors”. Since some librarians are travelers, then all travelers are counselors, that means that some librarians are counselors. *Yes. Yes, because there's a small amount of people of librarians that are travelers and if they're travelers, they're counselors. So some....so some librarians are travelers...or counselors rather.* “Some doctors are not dancers, all chefs are dancers”. This doesn't say there is a relation between doctors and chefs. Well you could have a doctor that's not a dancer but you can have a doctor that is a dancer, so you can't really conclude between chefs and doctors I don't think. Nothing. “Some athletes are nurses, no athletes are secretaries”. *So there's a small amount of athletes that are nurses, so we can split them up between nurses and non-nurses, but no athletes are secretaries. So....but that does not....that doesn't say that nurses can't be secretaries unless they absolutely have no time to do both.* So some athletes are nurses....no athletes are secretaries....I don't think you can conclude anything from this. Some athletes are nurses, no athletes are secretaries. Yeah, I think that's nothing.

The italicized portions of the interview above illustrate A.L.'s attempt to negotiate the sets of individuals that appear to add complexity to the task. A.L. repeats the premises and expresses what the quantifiers do and do not mean. The only difference between A.L.'s reasoning for both tasks is the complexity added by the quantifiers or sets of individuals in the categorical reasoning task. The need to consider sets of individuals represents a source of difficulty that is missing from conditional syllogisms. This is one possible reason for students' reduced performance on categorical syllogisms; categorical syllogisms contain an additional source of complexity in comparison to conditional syllogisms.

Moreover, notice how A.L. uses the phrases “does not say anything between them” and “doesn't say there's a relation” to signal the importance of a connection or link between premises as he reasons through both the conditional and categorical syllogisms (underlined portions in the

report above). The goal to establish a link between premises was achieved by students using a variety of strategies, which were reported primarily in the retrospective portion of the interviews (see Table 2). The concurrent reports were generally not useful in establishing unequivocally the strategies students used to reason through the syllogisms because most students did not mention their strategies explicitly. During the concurrent interviews at time 1, three students drew Venn-like diagrams for solving categorical syllogisms (only one of the students used a Venn-like diagram for solving conditional syllogisms). At time 2, two students used Venn-like diagrams for solving categorical syllogisms and two students wrote out symbolic-rules for solving conditional syllogisms. However, these strategies were observed on students' papers and were generally not expressed by students in their concurrent reports. The one exception was students' reported use of options in the SR format. The use of options was evident because students read the options aloud and thought aloud about them as they determined whether they led to a correct answer.

The concurrent reports at times 1 and 2 did also reveal students' systematic repetition of the syllogism premises as they reasoned through the syllogism (see A.L.'s report described previously). The repetition of the premises could be viewed as a basic strategy in so far as the premises are statements or rules that must be manipulated in order to generate a conclusion. Conversely, the repetition could be viewed as students being overly focused on the specific premises or *surface features of the tasks* and failing to generate an integrated plan of execution for actually solving the syllogisms.

Students' performance on categorical syllogisms was consistently better in the selected-response format (see Table 1). Identifying the link between sets of individuals may be alleviated by the selected-response format because this format presents an exhaustive set of options illustrating the possible connections between sets of individuals. The student does not have to generate the possible connections; all the student has to do is to evaluate the set of possible connections presented in the options because the correct answer is known to be among the options. In contrast, students' performance on conditional syllogisms was consistently better for the constructed-response format (see Table 1). One reason why conditional syllogistic performance may be better in the constructed-response format may be because students can construct their answers without the uncertainty that options can generate (Sloman, 1994). When individuals are more confident in their knowledge, the options illustrated in selected-response format can add uncertainty to an already formulated idea (Sloman, 1994). It is normally assumed that recognition is superior to recall because the task of recognizing and selecting a response from a list of alternatives activates more memory sources than the task of recalling and constructing a response (Anderson, 1990; Noice & Noice, 2002; Tulving & Thompson, 1973; Watkins & Tulving, 1975; Sloman, 1994). However, Tulving and Thompson (1973), Watkins and Tulving (1975), and more recently, Noice and Noice (2002) indicate that recall can be superior to recognition when the student is very confident about his or her performance or when the material is very well learned.

The third theme in students' concurrent reports was the *consistency* in their reasoning goals across time 1 and time 2; fourteen out of 16 students explicitly strived to establish a link or connection between premises. Consider A.L.'s report at time 2; it is very similar to his concurrent report at time 1 (described previously):

A.L.: Some veterinarians are not runners and some housekeepers are runners, so in a group of veterinarians, some are not runners. So these people in this little circle here are not runners and then some housekeepers are runners. Doesn't...hm.....well it ___ some people that aren't runners....I don't think you can conclude anything from this since some housekeepers, they're not tying it between the housekeepers and the veterinarians....just because some are and some aren't, that doesn't mean that they both are. Some librarians are travelers, all travelers are counselors, so for every

traveler there are counselors, and some librarians are travelers, that means some librarians are counselors. I didn't even spell that right.

I: That's ok.

A.L.: Some doctors are not dancers, all chefs are dancers. So if all chefs are dancers, but it doesn't mean that all dancers are chefs and if some doctors are not dancers.....then if all chefs.....for every chef they're a dancer and then some doctors are not dancers, so then some doctors are not chefs....are not chefs....because if a doctor....some doctors may be dancers....no, I'm going to retract that statement. Some doctors are not dancers, all chefs are dancers. So if a doctor is a dancer, that doesn't mean he's a chef, so nothing can be concluded. And some athletes are nurses, no athletes are secretaries. So if you're an athlete and a nurse, then you can't be a secretary because....well if you're an athlete, you can't be a secretary and since some athletes are nurses, that means that some nurses aren't secretaries....aren't secretaries. Since if you were an athlete and a nurse, you couldn't be a secretary. Ok.

Given that students' overall reasoning goals between time 1 and time 2 were correctly focused on finding a link or connection between premises, it is surprising that their performance did not improve with direct instruction. The course in logic provided them with the basic vocabulary and strategies in conditional (sentential) logic and categorical (predicate) logic to achieve their reasoning goals. Moreover, at time 2, 10 out of 16 students reported expected grades in their symbolic logic class that were above average, and 6 out of 16 students reported being at the top of their class. Thus, it can be assumed that the sample of students was performing well in the course.

The syllogisms that were used in the present study were also well matched to the objectives of the logic course and simpler in overall presentation than the material and homework assignments covered in the course. Thus transfer of training was thought to be quite good (Sternberg & Ben-Zeev, 2001). Of course, our tasks could not bear *exact* resemblance to the materials and homework assignments covered in the course because then we risked measuring recall and not reasoning (see Barnett & Koslowski, 2002). An assumption with most assessments is that the assessment tasks will be drawn randomly from the same population of tasks as those used during training. Therefore, students should be able to apply acquired knowledge to the assessment tasks even if the assessment tasks do not bear exact resemblance to the surface features of the course tasks used to learn the concepts. In fact, this is one reason why some cognitive psychologists refuse to include students with any logic training in their studies of reasoning; the expectation is that students with any training in logic will be at ceiling levels on categorical and conditional syllogisms (see Johnson-Laird & Bara, 1984; Schaeken, De Vooght, Vandierendonck, & d'Ydewalle, 2000; Stanovich, 1999). Our study shows that this is clearly not the case.

In sum, the concurrent reports revealed that categorical and conditional syllogisms were not ambiguous to students. Students were not confused by the words of the tasks. None of the students expressed frustration or confusion as they solved the syllogisms. The concurrent reports did suggest that categorical syllogisms are more challenging than conditional syllogisms at time 1 but students were still performing modestly on categorical syllogisms at time 2. If the task features were not confusing, then what could possibly be thwarting students' performance—especially students who reported to be performing well in their course? The retrospective portion of the interviews were reviewed to determine students' overall comprehension of the tasks, including sources of uncertainty, and their reported reasoning strategies.

Retrospective reports at Times 1 and 2. For the first question (i.e., which problem or problems did you find most difficult and why?), over 85% of students across times 1 and 2 identified

correctly those categorical and conditional tasks, which are considered most difficult. Likewise, students had no trouble identifying the easy categorical or conditional tasks. One of the dominant sources of task difficulty was attributed to the absence of a clear connection or pathway between premises. Over 90% of students identified the need to draw a link between premises as a source of difficulty. Consider the report from D.W. who reported being at the top of his class at time 2. D.W. expresses correctly the order of difficulty of the categorical syllogisms:

D.W.: The fourth problem – some athletes are nurses and no athletes are secretaries would be the....was what I found the most difficult. In both the first and the third, we have two separate groups in the first section – veterinarians and housekeepers and we talk about them belonging to the group of runners – as a subset type thing. And then in the third we talk about doctors and chefs and them being dancers. You can see....it's a little easier to picture two different groups belonging to something like you look at a group of boys and a group of girls with some common thing and you're thinking of similar situations whereas in the fourth, you start with one group and you say some of them are this and none of them are something else. So then you're saying is it possible for the group that some of them....without being to weird, the group that some of them belong to ie, the nurses, and some of them being the athletes, say is it possible for these nurses to be secretaries or must they be or can they not be. But then you realize that there are other athletes that are...sorry, other nurses that aren't athletes and could therefore belong to the secretaries group. *So it sort of has a couple of almost hidden scenarios I guess you know? You don't just see right away that the....I mean the other one – the "some" always sort of catches you, but here it's the negation like saying no athletes can be secretaries, so you're like oh, well since some athletes are nurses, that means that the nurses are one step back – were athletes. But that's not true for all nurses.*

D.W. now expresses the strategy he used to answer the categorical syllogisms at time 2:

D.W.: I think here is definitely sort of the idea of subsets and again, to mention the course or whatever, I think you get more adept at quickly recognizing the different sort of possible scenarios or sort of....and like immediately being able to recognize what the different groups are ie, like for the fourth one realizing that there are nurses there that weren't athletes and could therefore belong to the secretary group. Sort of seeing that as opposed to....like sort of judging the argument rather than just sort of following it as if it were true and then getting sort of caught up in it all.

There is a subtle contradiction in D.W.'s reports. To the question of task difficulty, D.W. correctly identifies the ratios or sets as difficult. To the question of strategy, however, D.W. reports quickly recognizing subsets. The contradiction is that while D.W. recognizes the difficulty of the task (see italics in report above), he spends little time spent trying to understand its important features. From D.W.'s responses to these questions, it appears that D.W. is failing to use his recognition of task difficulty to modify the time spent understanding the task fully, including most importantly the quantifiers in the categorical syllogisms. D.W. describes his grouping heuristic as "quickly recognizing the different sort of possible scenarios or sort of....and like immediately being able to recognize what the different groups are..." The problem with this approach is that D.W. is racing past a critical aspect of categorical reasoning—understanding the nature of the sets of individuals to

be considered in the reasoning process. Therefore, the grouping heuristic is quickly applied without fully understanding the features of the task that make this strategy successful.

D.W. correctly answered the categorical syllogisms 50% of the time at time 2. By examining D.W.'s retrospective report, we can understand why a student might be performing at 50% on the assessment task. The retrospective report informs us that D.W. is not spending enough time comprehending an important feature of the task (i.e., the meaning of the quantifiers). Going back to D.W.'s concurrent report shown below further supports the lack of time spent on the quantifiers. His conclusions are generated quickly even to the last, most difficult syllogism (i.e., Some doctors are not dancers; all chefs are dancers).

D.W.: “Some veterinarians are not runners, some housekeepers are runners”. A: All veterinarians are housekeepers....some veterinarians are not runners, some housekeepers are runners.....this doesn't really show any relationship between someone being a housekeeper and being a veterinarian, so I would conclude nothing. There's no transition between going from something to runner then runner to something. So second: “Some librarians are travelers, all travelers are counselors”. We can therefore conclude that any librarian that may be a traveler must be a counselor and therefore some librarians are counselors and that would be B. “Some doctors are not dancers, all chefs are dancers”. This again does not really show a relationship between doctors and chefs – merely between doctors and chefs and their love of dancing, so we would conclude nothing. “Some athletes are nurses, no athletes are secretaries”. Um.....if someone is an athlete they cannot be a secretary, however, only some athletes are nurses so it's possible to not be an athlete and still be a nurse, so I don't believe this concludes anything about the relationship between nurses and secretaries. Some athletes are nurses.....no athletes are secretaries.....because there could be other nurses that have nothing to do with athletes. So yeah, I would conclude nothing about the relationship between nurses and secretaries.

Table 2
Frequency of Strategies Reported for Categorical and Conditional Syllogisms by Time in Course and Item Format

	Time 1				Time 2				Total
	Categorical		Conditional		Categorical		Conditional		
	CR	SR	CR	SR	CR	SR	CR	SR	
Statements	1	4	2	1	1	1	2		12
Rules	4	1	5	4	3	1	4	4	26
Pictures	3	2	4	3	7	6	4	2	31
Heuristics		4		1		2		2	9
Total	8	11	11	9	11	10	10	8	

Note. Column frequencies do not always add up to 8 because some students reported multiple strategies

Other students exhibited similar problems fully comprehending the nature of the quantifiers in the categorical syllogisms, irrespective of their reported grade in the logic course. Consider K.R.'s comments at time 2 (K.R. reported being below average in the course):

K.R.: I would say although I should probably say I used rules because this class is almost over, but mostly I used a mental picture, sort of a circle representing all of something and then an arrow if it also implies another group and whether those two would somehow overlap or be the exact same thing. So I think that's why it gets confusing with the sentences which I at first feel have no connection to each other because it's just to separate circles with no arrows, and implies nothing.

K.R. attempted to use a reasonable strategy (Venn-like diagrams or pictures) but her efforts were unsuccessful because she quickly assumed the categorical premises implied no links. Little time was spent understanding that links could be drawn between premises and conclusions. Now consider M.H.'s concurrent and retrospective reports at time 2. First, from M.H.'s concurrent report, we see that M.H. reflects on the meaning of the quantifiers as she considers what they imply about the premises:

M.H.: OK. So for the first one *“Some veterinarians are not runners, some housekeepers are runners”*. And then from that I guess you could say that it wouldn't necessarily be true that any veterinarians would be housekeepers, as it's “some” in both of them and some veterinarians are not runners could imply that some of them were runners, but not necessarily because some could mean all. And the same goes for housekeepers being runners. So I think that the answer would be E. “Some librarians are travelers, all travelers are counselors.” So since it's all travelers are counselors, you can put counselors in where travelers are and then you could say that some librarians would be counselors as well. So B. *“Some doctors are not dancers and all chefs are dancers”*. So if all chefs are dancers, then the same as number two, you can take chefs and put it where dancers is in the first one and have some doctors are not chefs. So then that would be C. And for the last one: *“Some athletes are nurses, no athletes are secretaries”*. So it's some athletes are nurses and if they are athletes, then they can't be secretaries, then it doesn't necessarily mean that anything between the secretaries and the nurses because they're both...it could be...no, it does have to mean that some of the nurses are not secretaries because none of the athletes are secretaries and there's at least one athlete that is a nurse. So then it would have to be that some nurses are not secretaries. So that's C.

The critical portion of M.H.'s report is italicized. During her reasoning through the most difficult syllogisms, she considers the quantifiers by reporting what the quantifiers can and cannot imply. She also applies a heuristic called a *substitution strategy*, which is useful with difficult syllogisms that contain quantifiers. The substitution strategy involves assigning letters or symbols to sets in order to make the connection between sets of individuals more transparent. She alludes to the substitution strategy in her retrospective report:

M.H.: I guess I was kind of thinking is not of the things...like for the first one, I wasn't thinking of veterinarians as actual veterinarians, just think of it as like veterinarians as some A are not runners or like another symbol for it instead of thinking of it as actual things. It makes more sense if you don't think of actual people being the veterinarians.

I: So what would you call that then? Would you call that....did you imagine anything as you were solving the problems or did you use a rule or did you use a mixture of imagining and rules?

M.H.: Well I guess it's kind of imagining just because you have to change the things in your mind and you're not like rewriting them all out, but I think it's also....I guess it's a mix because it's kind of a rule too because you could always do it the same way, you don't have to change it for each individual one, you just substitute the things.

M.H. scored 100% on the categorical syllogisms and also reported being at the top of her logic class. Another student, this time male, who also scored 100% on categorical syllogisms, reported using the options directly in his reasoning because of the difficulty of the quantifiers:

H.M: I think that I was trying to...I was using the same strategy I described before – where I would after reading the statements I would try and consider what was true of...or what was the conclusions you could draw from the statements. But I think on some of those that I had a bit more difficulty, I started looking at the multiple choice options and thinking about could they be true given the statements about...and then applying each one, determining which ones were true or which ones were true and which ones were false.

Although over 80% of students at time 1 and over 85% of students at time 2 identified categorical syllogisms as the more difficult of the two tasks, the majority of students did not use this information about task complexity to solve categorical syllogisms more slowly or carefully. C.F. elaborates further on the complexity of the categorical syllogisms:

C.F.: Um.....well in this one you're dealing with one person, so just in the place and they're either in that place or not. So it's actually maybe I'm using mental pictures for these too. But it's easier to just sort of go o.k, he's there or she's there. For this one you're starting to talk about ratios and groups and for me, I don't know, that's how I saw it. And determining who was part of what and the language was more....because there was more connectives and things. So I thought it was a little more complicated.

The retrospective reports suggested that students are aware of the difficulty of the tasks they solve. This is good news because it suggests that the difficulty level of higher-level reasoning items is aligned with students' interpretations of difficulty. Moreover, students were not confused by the task or by the instructions. This is also good news. Students understood, at least at a surface level, what was required of them as they progressed through the tasks. Nevertheless the retrospective reports did reveal a source of misalignment between students and tasks. The source of this misalignment, however, did not originate from incomplete task instructions or ambiguous words. The source of the misalignment originated from students who did not seem to devote enough time to fully comprehending important conceptual features of the tasks they were solving. A large majority of students, even at the end of their 12 week course in symbolic logic, were still uncertain about the meaning of quantifiers and also uncertain about the time required to represent quantifiers properly. It is only after appropriately representing the quantifiers that strategies can be used successfully (as did M.H. when she selected the substitution strategy to identify the link between sets). In sum, one reason why students in this study may not have improved substantially in their reasoning at time 2 is because they appeared to race past a critical stage in problem solving—the comprehension stage.

Although some of the students in this study reported being at the top of their class and may have learned strategies to solve categorical and conditional syllogisms, these strategies were misapplied in a testing situation that differed from their course context. Why would students not be able to apply strategies learned in one context to another similar context?

One hypothesis that is receiving increasingly greater support is that students are unable to transfer acquired knowledge because they lack skill in comprehending the important objectives of the tasks they solve (Kuhn, 2001; Leighton & Gokiert, 2005; Nickerson, 2004; Shaughnessy, 2004). Task comprehension is a metacognitive skill that successful problem solvers within a domain must master. Task comprehension does not mean simply understanding the words of the task, but rather, it means understanding its important conceptual features and generating an integrated representation of the task's intended objective. Expert problem solvers spend most of their time negotiating what is a proper task interpretation (Ericsson & Charness, 1993). Only once the task is well understood, can the successful problem solver select the pathway for problem solution. As evidenced by concurrent reports and retrospective reports, students in our study identified proper reasoning goals, and *attempted* to use proper reasoning strategies, but their performance suffered because a majority failed to fully comprehend the important features of the tasks they were asked to solve. Soon after reading a syllogism, students started evaluating response options in the selected-response condition or started to respond in the constructed-response condition.

General Discussion

The purpose of the present research was to (a) investigate the alignment of categorical and conditional syllogisms for measuring higher-level reasoning in students and (b) demonstrate the feasibility of using both concurrent and retrospective think-aloud methods for collecting evidence of students' cognitive processing on reasoning test items (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995; Taylor & Dionne, 2000). We have found that categorical and conditional syllogisms evoke the proper problem solving *goals* in students—almost all students (14 out of 16) expressed the need to establish a link or connection between premise sets in order to generate a solution. The identification of this link indicates that students understand the surface objective of syllogisms. Implication is basic to generating necessary conclusions to logical tasks and is an essential aspect of logical and higher-level reasoning according to cognitive psychologists (see Johnson-Laird & Bara, 1984; Powers & Dwyer, 2003; Stanovich, 1999). However, understanding that syllogisms involve identifying a link between premises and conclusion is not enough. The actual task features (e.g., quantifiers) are also important and must be negotiated carefully (and not at a surface level) in order to generate a correct representation of the task's objective. Students who identified the importance of the link but missed the quantifiers in categorical syllogisms did not perform well because they spent little time decoding important task features.

In other words, the one source of misalignment we found rests less with the tasks and more with the students. We found evidence that students were taking little if any time to properly comprehend the nature of the tasks—especially the categorical syllogisms that contained quantifiers and thus contained more complexity than conditional syllogisms. Generating a proper mental representation of a task is a necessary step before a strategy is selected and properly executed. Although over 85% of students correctly identified categorical syllogisms as the more difficult set of tasks, there was no evidence in students' concurrent reports or retrospective reports that they adjusted their problem solving to deal with the task complexity. Only two students out of 16 scored 100% on the categorical syllogisms at time 2. The two students used heuristics to deal with the quantifiers systematically—one student claimed to have used a substitution strategy and another student used the SR options.

That students themselves may be causes of misalignment seems unlikely because test items are normally assumed to contain the sources of error (see Ferrara et al., 2003, 2004; Haladyna & Downing, 2004; Leighton, 2004; Leighton & Sternberg, 2003; Roberts & Newton, 2005). However, student-based misalignment is not as uncommon as we think (Shaugnessy, 2004). In recent reports focused on students' metacognitive skills, researchers have found that students who are unsuccessful across tasks and domains have poor skills at comprehending tasks and problem-solving situations; their comprehension is rooted in surface features rather than broader concepts (Kuhn, 2001; Shaugnessy, 2004). In other words, students who perform well within highly specific domains and for certain tasks alone, tend to have skills that are not conceptually based but more situation-based. Thus, if the situation is changed, the students' performance drops. This is a serious distinction in students' performance—whether students perform well based on either conceptual understanding or surface understanding. A student who performs well on a standardized test *expected* to measure high-level reasoning may be performing well because he or she has studied a specific class of items so well as to know how its surface features relate to an item responses. However, if the scope of the inference generated for a student applies only to a specific class of items, then the inference is weak. We would like to generate inferences from test performance that cut across similar situations; that is, inferences with a broader range. This is similar to the challenge posed by test-wiseness (see Rogers & Bateson, 1991). We must develop assessments that distinguish students who perform well because they understand concepts and have mastered skills from those who have become adept at associating specific responses with surface features of tasks. The use of concurrent and retrospective interview procedures may provide a method to begin to understand not only how students solve test items but also how they approach them conceptually.

References

- Aikenhead, G.S. (1988). An analysis of four ways of assessing student beliefs about STS topics. *Journal of Research in Science Teaching*, 25, 607-629.
- Aikenhead, G.S., Fleming, R.W., & Ryan, A.G. (1987). High school graduates' beliefs about science-technology-society. I. Methods and issues in monitoring student views. *Science Education*, 71, 145-161.
- Aikenhead, G.S., & Ryan, A. G. (1992). The development of a new instrument: "Views on Science-Technology-Society" (VOSTS). *Science Education*, 76, 477-491.
- Anderson, J. R. (1990). *Cognitive psychology and its implications* (3rd ed.). NY: W. H. Freeman and Company.
- Barnett, S.M., & Koslowski, B. (2002). Adaptive expertise: Effects of type of experience and the level of theoretical understanding it generates. *Thinking and Reasoning*, 8, 237-267.
- Brunkhorst, H. (August, 1987). *A comparison of student/ teacher positions on selected science-technology-society topics: A preliminary study*. Paper presented at the 4th International Symposium on World Trends in Science and Technology Education, Kiel, West Germany.
- Cheng, P., Holyoak, K. J., Nisbett, R. E., Oliver, L. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, 18, 293-328.
- Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NJ: Lawrence Erlbaum.
- Embretson, S. & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38, 343-368.
- Ercikan, K., Law, D., Arim, R., Domene, J., Lacroix, S., & Gagnon, F. (April, 2004). *Identifying Sources of DIF Using Think-Aloud Protocols: Comparing Thought Processes of Examinees Taking Tests in English versus in French*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis*. Cambridge, MA: The MIT Press.
- Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, 49, 725-747.
- Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition*, 77, 1-79.
- Ferrara, S., Duncan, T.G., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., Mattessich, A., Rogers, A., & Westphalen, K. (April, 2004). *Examining test score validity by examining item construct validity*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Ferrara, S., Duncan, T., Perie, M., Freed, R., McGivern, J., & Chilukuri, R. (April, 2003). Item construct validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Giroto, V. (2004). Task understanding. In J. P. Leighton and R. J. Sternberg (Eds.), *Nature of reasoning* (pp. 103-128). NY: Cambridge University Press.
- Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, Spring, 17-27.
- Irvine, S. H. & Kyllonen, P. C. (Eds.), (2002). *Item generation for test development*. NJ: Lawrence Erlbaum.
- Johnson-Laird, P.N. (1999) Reasoning: formal rules vs. mental models. In Sternberg, R.J. (Ed.) *Conceptual issues in psychology*. Cambridge, MA: MIT Press.

- Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, 16, 1-61.
- Klenck, V. (2001). *Understanding symbolic logic 4th edition*. Prentice Hall.
- Kuhn, D. (2001). How do people know? *Psychological Science*, 12, 1-8.
- Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday life events. *American Psychologist*, 43, 431-443.
- Leighton, J. P. (under review). *Teaching and Assessing Deductive Reasoning Skills*.
- Leighton, J. P. (2004). Avoiding Misconceptions, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice, Winter*, 1-10.
- Leighton, J.P. & Gokiert, R.J. (2005, April). *The Cognitive Effects of Test Item Features: Informing Item Generation by Identifying Construct Irrelevant Variance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Montreal, Quebec, Canada.
- Leighton, J. P., & Sternberg, R. J. (2003). Reasoning and problem solving. In A. F. Healy & R. W. Proctor (Volume Eds.), *Experimental Psychology* (pp. 623-648). Volume 4 in I. B. Weiner (Editor-in-Chief) *Handbook of psychology*. New York: Wiley.
- Morris, M. W., & Nisbett, R. E. (1993). Tools of the trade: Deductive schemas taught in psychology and philosophy. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 228-256). Hillsdale, NJ: Erlbaum.
- Munby, H. (1982). The place of teachers' beliefs in research on teacher thinking and decision-making, and alternative methodology. *Instructional Science*, 11, 201-225.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Nickerson, R. (2004). Teaching reasoning. In J. P. Leighton and R. J. Sternberg (Eds.), *Nature of Reasoning* (pp. 410-442). NY: Cambridge University Press.
- Noice, T., & Noice, H. (2002). Very long-term recall and recognition of well-learned material. *Applied Cognitive Psychology*, 16, 259-272.
- Powers, D. E., & Dwyer, C. A. (May, 2003). *Toward specifying a construct of reasoning*. Research Memorandum, Educational Testing Service, Research and Development Division, Princeton, NJ.
- Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum.
- Roberts, M. J., & Newton, E. J. (2005). Strategy usage in a simple reasoning task: Overview and implications. In M.J. Roberts and E.J. Newton (Eds.), *Methods of Thought: Individual differences in reasoning strategies*. Psychology Press.
- Rogers, W. T., & Bateson, D. J. (1991). Verification of a model of test-taking behavior of high school seniors. *Journal of Experimental Education*, 59, 331-350.
- Schaeken, W., De Vooght, G., Vandierendonck, A., & d'Ydewalle, G. (Eds.). (2000). *Deductive reasoning strategies*. Mahwah, NJ: Erlbaum.
- Shaughnessy, M. (2004). An interview with Deanna Kuhn. *Educational Psychology Review*, 16, 267-282.
- Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition*, 52, 1-21.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Standards for Educational and Psychological Testing. (1999). *Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education*.
- Sternberg, R. J. & Ben-Zeev, T. (2001). *Complex cognition. The psychology of human thought*. Oxford University Press.

- Taylor, K. L., & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413-425.
- Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 359-380.
- Watkins, M., & Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General, 104*, 5-29.
- Yarroch, W. L., (March, 1986). Content validity of the 1985 Michigan department of education pilot science examination. Paper presented at the annual meeting of the National Association for Research IN Science Teaching, San Francisco, CA.