**Design of the study.**

Introduction.

My goal was to use the proposed theoretical "BARK" framework to gain a new insight into the processes of solving *non-trivial* physics tasks – that is, physics problems. To achieve my goal, I have designed the study with an eye on the inconsistencies found in prior studies. The study was novel in several distinct – but intertwined – ways:

- **Theoretical**: the proposed BARK dichotomy, while firmly grounded in the results of existing studies, combines the previously suggested cognitive components of problem solving processes in a new way, possibly more productive for research and instructional purposes;
- **Methodological**: based on the review of the literature, several important design features have been implemented; to my knowledge, no problem-solving study has ever combined them all – and I consider each of them crucial in achieving the desired results.
- **Logistical**: having been granted access to a new Internet-based platform, CyberTutor, which combines advanced ways of delivering tasks and collecting student feedback, allowed me to satisfy my methodological requirements and obtain the type of data that have not been available to researchers before.

Research sample.

The direct participants in the study were high-school AP Physics C[1] students. More than 140 students participated in the study; 103 of them submitted their background questionnaires and completed all or almost all tasks; the data collected from these students were used in subsequent analysis. In addition to the direct participants, I must credit about 250 MIT freshmen and their instructors who, in the course of their studies, solved some of the tasks and provided valuable feedback. That feedback, collected as part of the task-validity study, helped correct glitches, eliminate ambiguities and determine the relative difficulty of the tasks before they were offered to the direct participants.
I selected the AP Physics students as subjects for several reasons:

- At the core of my study is the understanding that a given task may be considered either a *problem* or an *exercise* (and require different sets of cognitive skills for solving it) depending on the background of the solver. Therefore, *I had to have the participants with very similar backgrounds in physics* in order to be able to assign the same set of tasks to all participants and observe the differences in the *problem*-solving performance of less successful and more successful problem solvers. While it is, of course, impossible to achieve perfect background similarity, all AP students follow the national AP curriculum and are a *nearly-idea*l population in that regard;
- *AP Physics C is the most challenging widely offered high-school physics course* and is arguably, the *only* course where meaningful problem solving takes place on a regular basis; besides, the national distribution of the AP Examination grades indicates a wide variability of their problem-solving skills; I could reasonably expect similar variability among my participants;
- *very few studies had explored the problem-solving abilities of high-school students*. While it is of great interest to explore the minds of high-school students, most of the research, according to the literature, has concentrated on college-level students.

Logistical reasons appear to have been the main reason for the lack of high-school-based studies: to construct a large high-school sample (for instance, of the size that I used), one must work with several different classes, maintain communications with several high-school teachers and find a way to administer the tasks and collect the feedback in a consistent manner. Besides, the students are overwhelmingly minors, and obtaining parental consent is an extra barrier to be overcome before the study can begin. On the other hand, an introductory physics class in colleges can easily involve one hundred students or more, and no parental consent is required. Many researchers have been content to use their own students (as I mentioned, that fact alone raises some validity concerns in my mind) or the students of their fellow professors.

---

[1] AP (Advanced Placement) Physics C is a calculus-based, college-level course; it is the most advanced physics course available to most high-school students in the U. S.

I was very interested, however, in studying high-school students; my position as a high-school teacher and the availability of CyberTutor allowed me, I believe, to overcome most of the hurdles that had precluded the authors of large-scale problem-solving studies from using high-school student participants.
Here is how the sample was constructed. In September-October, e-mail letters were sent out to about sixty AP Physics teachers nationwide, asking them to have their students participate in the study. The teachers and the students were told that I would observe the behaviors during the problem-solving processes; however, the details of the BARK theory and the detailed design of the study were not made available to the teachers and the participants. Great care was taken in making the letter sound "non-threatening" to prevent the self-selection of only the top teachers. [2]
Finding the volunteers was not easy at all: most of the teachers were not very eager to take on another responsibility, and I did not have any means of compensating them or their students for participation[3]. However, ten teachers agreed to participate, of which nine actually did. These nine teachers were mailed the consent forms, which were to be signed by both the student and the parent/guardian. Besides, the students were given the background questionnaires, to collect the information possibly relevant to their problem-solving abilities. The design of the questionnaires (see Appendix A) was based on the available literature (Chi *et al*, 1989; Hammer, 1989; Harpole & Gifford, 1986; Sadler & Tai, 1997, 2001). The consent forms and the background questionnaires were distributed, collected and then mailed directly to me by the teachers. At my suggestion, most teachers made my tasks a part of the regular homework, and, overall, more than 140 students attempted at least some of the tasks. However, in my research, I used only the data provided by the students who submitted their consent forms.
After the CyberTutor tasks were completed, student interviews were conducted. For the interviews, I selected eight participants, all of whom had consented to being interviewed. Six participants were male and two were female; the participants came from two different local schools. To observe a wider spectrum of problem-solving behaviors, I selected the students who had shown different degrees of success solving the CyberTutor tasks[4].

<u>The tasks</u>.

Besides the background information provided by the students, the study included the data collected from the students who solved the Mechanics tasks administered via Internet by CyberTutor and the interview protocols from a sub-sample of the participants. Three major considerations made me choose Mechanics as the focus of the tasks offered to the students to solve. First, mechanics is the area where most of the problem-solving research (including my own pilot study – see Korsunsky & Record, 2000) has been conducted. Second, the students taking AP Physics C study Mechanics in the first semester of the school year (typically finishing in December-January). By November-December, the students accumulate enough knowledge to solve fairly complex mechanics problems. Since the students then take the AP test in May, the teachers always try to maintain the working knowledge of mechanics throughout the year, making them comfortable with multi-concept mechanics tasks throughout the second semester of the academic year[5]. The third consideration is that, having written thousands of physics tasks for several major publishers, I have considerable experience in creating problems of different styles and various degrees of difficulty – more in Mechanics than in other areas.
The combination of these circumstances made Mechanics a perfect domain for my tasks.
The CyberTutor tasks included a 28-item multiple-choice Mechanics Diagnostic Test and five open-response problems.

---

[2] See Appendix B for the sample of the recruitment letter.
[3] Eventually, I ended up sending small checks, drawn from my personal funds, to the teachers. I asked them to buy ice cream for everyone involved; in two local schools, I hosted the ice cream parties in person. This largely symbolic way of compensation had been – of course – cleared in advance with Harvard Institutional Review Board, as the participants were to be subjected to large amounts of the high-fat, high-sugar, high-calorie food.
[4] The weakest interviewees had not solved any CyberTutor problems; the strongest solved all five.
[5] Some students in my study studied only Mechanics throughout the year; they were administered the tasks at a later time in the year, to ensure a similar level of competence.

Mechanics Diagnostic Test (see Appendix C) was designed by me especially for this study and included a mix of strictly qualitative questions as well as ones with a quantitative component present[6]. My intention was to combine 14 questions that probed the knowledge of the basics, or *rigid knowledge* (R-questions) and 14 bisociation-probing questions (B-questions). In B-questions, the concept or concepts relevant to answering the question was less than obvious, based on the background of a typical AP Physics C student. Each question had four answer options of which one was correct. The questions were designed and pre-classified by me as being of R-type and B-type by me beforehand; they were then validated as described in the **Validity Issues** section of this chapter.

The questions assumed prior knowledge of Kinematics, Newton's laws, Gravity, Rotation of a Point Object, Conservation of Energy and Conservation of Momentum – topics that are extensively covered in the AP Physics C curriculum. For each topic, there was an approximately equal number of R- and B-questions.

Five open-response Mechanics problems (see Appendix D) were also specifically designed by me for this study, although all of these tasks were derived from the problems featured in *Challenging Problems for Physics* (Korsunsky, 1995a). Each problem required combining multiple concepts in order to be solved; all tasks were quite challenging, in the opinion of the experts and according to preliminary testing with MIT freshmen.

Each problem contained several (6-8) hints. The possibility of administering hints-on-demand is a distinct feature of CyberTutor[7], which allowed me to probe the nature of student difficulties in a relatively direct way. The assumption was that the key to the nature of the hurdle that one faces while solving a problem lies in a hint that helps the student to clear that hurdle. I find this way of exploring student difficulties a good complement to the information that can be elicited from the student interviews.

I chose quantitative problems to probe students' problem-solving abilities for methodological/logistical reasons. First, when a problem has a definite correct numerical answer, there is no room for subjective judgment on the researcher's part (as in "So, did she or did she not solve this one?"). A second – equally important – reason is that the students needed to have a *clear real-time feedback* from the computer indicating whether or not the problem has been solved or that a hint may be needed. Facilitating such computer-generating feedback for the qualitative problems that require the narrative-type answers, was not feasible in this study – the technology is just not there yet. However, I would like to stress, again, that the proposed framework of the problem-solving processes can be reasonably applied to qualitative problems as well.

The hints were classified, invisibly to the students, as R-hints and B-hints. The R-hints (for Rigid knowledge) provided the information such as lists of equations and reminders about basic algorithms and the B-hints (for Bisociation) suggested what principles, approaches or ideas may be relevant to the solution. To discourage the students from using the hints too liberally, they were told that a small amount of credit would be deducted for each hint used; also, each hint had a title, which the student could see before opening the hint itself. Based on the titles, the student would only open the hints deemed useful at a particular point in solving the problem. I must say that, since the students were not aware of the specifics of the study, I took a lot of complaints that the hint titles did not make it clear what the hints were all about – however, if the titles were too informative, they would have served as hints themselves, and the goal was to avoid such situations.

In addition to the CyberTutor-based tasks, I selected four challenging problems from my book, *Challenging Problems for Physics* (Korsunsky, 1995a), to be used in student interviews. These problems were not supplied with any hints.

Data collection process.

The data in the study came from three major sources:

- student background questionnaires;
- the log of the student interactions with the CyberTutor tasks;

---

[6] I resent the "conceptual – quantitative" dichotomy, often found in literature. There is no reason a quantitative problem should be "concepts-free", of course. In my opinion, in a good quantitative problem, it is the *conceptual* component that presents the main challenge. For that reason, I am careful to use a "qualitative – quantitative" dichotomy.

[7] See Appendix E for the examples of the various screen views showing the hints to a problem.

- student think-aloud protocols collected during interviews.

*The background questionnaires* (see Appendix A) were collected along with the consent forms at the initial stage of the project. The data from the questionnaires were then tabulated and analyzed statistically along with the performance data collected from the CyberTutor tasks.

Administering the *CyberTutor tasks* began after the students had learned all the concepts involved in solving the tasks. Each student was given a password and a username to log on to CyberTutor. All work done by the students was unsupervised; however, several special precautions[8] were taken to ensure the highest possible level of compliance with the rules. No rigid time limits were set for the CyberTutor tasks; however, the instructions gave the *suggested* time for each task.
For the first CyberTutor assignment, *Mechanics Diagnostic Test*, the suggested time limit was one hour (all students finished in less time than that). The multiple-choice questions, with four options each, appeared on the computer screen one at a time, and, once the Submit button was pressed, the answer to a question could not be changed. The students, therefore, had only one attempt for each question.
*Five open-response problems* were administered in two installments: the first one included three problems and the second one included two. The suggested time limit for each problem was set at 40-50 minutes. In the on-screen instructions, the students were reminded about the importance of working alone, taking the assignment seriously and using the hints thoughtfully. The students were advised that they were given five attempts for each problem, in order to enable them to use hints and to reduce the influence of possible technical errors. Also, the instructions encouraged the students to submit a written comment describing the solution process after finishing the problem (in the context of the study, a "finished" problem was either "solved" or "given up on").
Once the students opened a problem[9], the problem statement, the question and the answer field were immediately visible. The student could submit an answer and get immediate (correct/incorrect) feedback. If the student had difficulties, an on-screen Hints button could be pressed, and the synopses of all available hints appeared. After reading the synopses, the student could click on the one that seemed potentially useful and open the hint itself. The answer could still be submitted at any point during the process. Also, the students were allowed to exit and re-enter the problem.
For each series of interactions with a problem, three outcomes were possible:

- problem solved correctly;
- problem not solved, all five attempts used;
- problem not solved, some attempts remain but the student gives up.

"Giving up" could be done by either exiting the problem or pressing the Give Up button. Either of these actions would take the student to the screen where the comments regarding the solution process could be entered, and the final exit from the problem commenced. The comment screen also appeared after the problem was solved or when all attempts had been used.
Notably, the time spent by the participants on a given task was not reported as data, even though CyberTutor kept a real-time log of all student interactions with the computer. The readers may recall that all tasks were done in the "homework regime", unsupervised, so the students often took breaks, tried the same problem over the course of several days, etc. Under these circumstances, reporting the time spent on each problem would have been meaningless; however, in future studies, where the participants will be solving tasks under supervision, the time spent solving a problem will be of interest indeed.
*Student interviews* commenced after the CyberTutor tasks had been completed, in April-May. The students were in their final stage of preparation for the AP Exam, and their knowledge of the material learned in the fall and winter did not seem to have faded. I selected eight participants: six male and two female, from two different classes. The format of the interviews was *researcher-cued think-aloud protocols*, which have long been used in problem-solving studies (Chi, Feltovich, & Glazer, 1981; Chi *et al*, 1989; Finegold & Mass, 1985). According to this method, the students were given a task to solve and were encouraged to think

---

[8] See the **Validity Issues** section.
[9] See Appendix E for examples of different windows visible to the students at different stages of solving the problem.

aloud as the problem-solving process unfolds; I sat nearby, occasionally asking clarifying questions but keeping the intervention at a minimum[10].

Each participant was interviewed twice. In each interview, two problems were offered, with a suggested time limit of 25 minutes each thus keeping the length of each interview under 45-50 min. (Miller, 1991). In the event, most students either solved or gave up on each problem within the set time limit, although occasionally I had to stop the students in order to proceed with the second problem. The audio recordings and the notes taken during the solving process were then collected and analyzed. Altogether, I collected 32 think-aloud protocols.

The overall procedure of data collection was rather complex – and the scale rather large – compared to most problem-solving studies. Using a large national sample, the multitude of original tasks and both qualitative and quantitative methods of analysis[11] made the study innovative – but how valid are the results? Fully understanding the importance of rigorously enforcing (and sometimes, I believe, exceeding) the established standards of validity makes me confident that the data that I collected were authentic and meaningful, and the methods of analysis appropriate. In the next section, I will discuss the specific measures taken at various stages of the project to ensure the validity of the study.

**Validity issues**.

<u>The sample</u>.

As mentioned earlier, I aimed to have a group of problem solvers with similar background in physics. In most other studies (de Jong & Fergusson-Hessler, 1986, 1991), this goal was achieved by simply recruiting all participants from a single college class. Such method of selection raises two concerns. First, the results, whatever they may be, may be skewed by the teaching style of the instructor teaching the course: the students may have been taught (or not taught) a particular approach or technique, etc. Second is the fact that the researchers are often the ones teaching the participants, which inevitably raises concerns about conflict of interest. Both of these issues are mitigated by choosing the students from nine classes nationwide. All students followed the nationally established AP curriculum, ensuring the similarity of the participants' background in physics. None of the classes was taught by me, and the participation was strictly voluntary.

My method of sample construction also ensured a relatively narrow range of general academic abilities, as measured by traditional indicators (course grades, PSAT and SAT scores, etc.). For the AP students, the range of these indicators tends to be quite narrow, on the high end[12]. Meanwhile, the students' problem-solving ability, as measured by the results of the national AP exam, varies widely (in 1997, about 24% of the exam-takers got the top grade (5), and about 29% got 1 or 2). This combination of factors provides for a good chance to observe a relatively wide range of physics-specific problem-solving abilities among the individuals who are, generally, good students. Still, I collected the information about the factors that are known to correlate with physics achievement and used it in data analysis.

Also, I attempted to have a gender-balanced sample, since gender is known to be one of the major factors associated with students' academic participation and success in mathematics and physics (Sadler & Tai, 2001; Strauss, 1988).

It would have been difficult for me to construct such a sample given serious difficulties with recruitment. However, by the grace of fate, my final sample contained 65 male and 38 female students. Thus, the proportion of females (37%) slightly exceeded the national percentage of enrollment in AP Physics C (27%)[13] and also allowed me to have a large enough number of females to conduct meaningful regression analysis. Other measures taken included recruiting female teachers (two females out of nine participants) and a gender-balanced group selected for the interviews (two females out of eight participants).

Finally, a few words about the "self-selection bias." As in *any* volunteer-based study, one can, technically, argue that the participants were somewhat different from the general population just by virtue of having agreed to participate. However, I took care to minimize any possible bias. The soliciting letters[14] stressed

---

[10] See Appendix G for examples of interview protocols.
[11] See Chapter 3, **Data Analysis**.
[12] See <u>www.collegeboard.com/ap/techman</u>.
[13] See <u>www.collegeboard.com/ap/techman</u>.
[14] See Appendix B.

the non-judgmental nature of the study, ensuring that even less experienced and less skillful teachers would be comfortable participating. Indeed, the teachers who agreed to participate ranged widely in their age and teaching experience. Each teacher also had an extensive phone interview with me before the study was commenced. The purpose was to address any possible concerns on the teachers' part and ensure that they are not in any way tempted to "show that their kids are the best." Also – perhaps even more importantly – the recruited teachers encouraged *all* their students to participate, stressing the non-judgmental character of the study. With rare exceptions, *all* students of the participating teachers signed the consent forms, thus ensuring that I would not be working only with the strongest – or the most adventurous – students in a given class. Indeed, the profile of my participants indicates that my sample, both demographically and academically,[15] closely reflected the population of AP Physics C students nationwide. Therefore, I am confident that the self-selection bias in the sample, even if present, was minimal – especially when contrasted with the majority of the studies in the field of problem-solving in which the researchers tended to use their own students. Overall, I believe that the validity of my results is not, in any significant way, diminished by the selection process.

The tasks.

With my extensive experience as a problem-writer, I was well positioned to create meaningful and valid problem-solving tasks. As I mentioned, I had published literally thousands of problems and multiple-choice questions for various purposes: end-of-chapter tasks of various difficulty, items for the state and national standardized tests, exam and practice problems for the US Physics Team, etc. Besides, some of my prior research focused on designing unbiased multiple-choice tasks (Korsunsky & Record, 2000). I can fairly consider myself an expert in the field, capable of creating challenging problems and the multiple-choice items probing either rigid knowledge (R-questions) or bisociation (B-questions).

Of course, I was not content with simply declaring myself an expert. Before being administered to high-school participants, the tasks underwent a rigorous review process: they were screened by the group of MIT physics instructors and by the participating physics teachers. Based on their feedback, I eliminated mistakes and ambiguities, adjusted the balance among the topics covered, replaced weaker items, etc.

The multiple-choice tasks were screened especially rigorously, in order to ensure that each item does, in fact, correspond to its designation as an R-question or a B-question. To achieve that, the items were administered to several groups of MIT freshmen; their responses were analyzed, and the items edited and/or replaced based on that analysis. Notably, CyberTutor allowed me to see not only the difficulty rating (percentage of correct answers) but also the percentage of students who chose each of the four answer options; thus, I was able to fine-tune the difficulty and the focus of the items by replacing or editing the answer options.

Task delivery.

Administering the tasks through the Internet also helped ensure the *uniformity* of the procedures and enhance the validity of the results. For instance, by using the options provided by the CyberTutor platform, I was able to decrease the possible effects of *inductive influence* of questions on each other.

First of all, CyberTutor made it impossible for the students to view a question once it was answered, thus preventing them from taking a clue from one question and using it to answer another. Of course, the students got no feedback on the correctness of their answers.

Also, the questions were administered *in a pre-set order*. The participants were divided into four roughly equal groups (for all students in a given class, the order had to be the same); each group was administered the question in a different order. Altogether, there were four such orders of questions:

- RBBT (rigid knowledge, bisociation – by topic);
- BRBT (bisociation, rigid knowledge – by topic);
- RABA (rigid knowledge – all; bisociation – all);
- BARA (bisociation – all; rigid knowledge – all).

---

[15] See Tables 2a, 2b in Chapter 3 for the descriptive characteristics of the sample. The relevant characteristics of the national sample can be found at http://apcentral.collegeboard.com/repository/01_national_8128.pdf and http://www.collegeboard.com/research/abstract/0,1273,3861,00.html (both sites viewed on Nov. 11, 2002).

None of these measures, to my knowledge, had been taken in the well-known and widely accepted Force Concept Inventory and Mechanics Baseline Test studies, which used the pencil-and-paper format of the tasks.
Another issue to be reckoned with was, of course, **compliance**. It is always an issue when the tasks are solved by unsupervised students. However, the lack of supervision was a trade-off for the scale of the study: about a hundred students, nationwide, spent about six hours each working on my tasks over the course of several months, far exceeding the numbers of "participant-hours" in the studies that involved actual problem-solving by the participants.
In general, AP students tend to be very strong academically, and their academic integrity is, probably, very high. More specifically, the participating teachers felt strongly that their own students, with few possible exceptions, were 100% honest, hard-working individuals. Also, the students who had volunteered for the study can be reasonably thought of as willing to do the tasks in a manner consistent with my instructions; no pressure whatsoever was made by the participating teachers to help recruit the students.[16]
Therefore, the very method of constructing the sample provided for a high compliance level.
Nevertheless, several additional steps were taken to further enhance the level of compliance:

- the stakes were set "not too high and not too low": the teachers told their students that their duty is to do their best, and that it was an honest effort, not the results, that counted. Most teachers treated these assignments as mandatory – but not graded – homework, often giving extra points for doing well but not penalizing for unsolved problems.
- each assignment was preceded by a set of instructions reminding the students to work alone, with no use of textbooks and notes, etc., to stay close to the suggested time limit, etc.
- the log of the student actions was analyzed, and the obvious cases of cheating[17] were eliminated from the data pool; there were very few such cases, involving only eight of the 103 participants.
- also, based on the log of student actions, I eliminated several cases of "pretending to solve" when the student would simply randomly open all hints and use up all submissions within a very short period of time. Such cases were also very few[18].

Overall, the task delivery procedure was thoroughly developed, extensively pilot-tested and, with the help of the participating teachers, successfully and consistently implemented.

<u>The interview process</u>.

The interview process was the least "innovative" part of the data collection process. I used the technique of *think-aloud protocols*, which has been well-developed and frequently used in problem-solving studies (Chi, Feltovich, & Glazer, 1981; Chi *et al*, 1989; Finegold & Mass, 1985). In this format, I gave the participants the tasks to solve and encouraged them to talk themselves through the solution process, commenting on their actions and thoughts. After that, I just listened, keeping the intervention to a minimum: usually, I would only ask a clarifying question if I could not understand what a student meant by a particular phrase; also, when a student kept silent for too long, I would gently prompt the student to say something: "So, what are your thoughts right now?" Such interview format kept to a minimum a possibility that students' thoughts may be influenced by the researcher. Several other measures were taken to ensure the participants' comfort, sincerity and high performance during the interviews:

- all participants were volunteers who affirmed their willingness to be interviewed after they had agreed to participate in the study;
- the participants were both male and female, with a wide range of problem-solving abilities; thus, the interview sample was reasonably representative of the overall sample;
- none of the participants was my current or former student;
- before the interviews, I advised each participant that the problems are very hard, and that I look to observe the process of solving rather than judge their ability to solve problems;

[16] which is why it was so hard to find enough participants.
[17] For instance, submitting the correct answer within one minute of the first viewing of the problem.
[18] Importantly, none of the statistical outcomes changed noticeably after the data set was cleaned from the actions of "cheaters" and "pretenders."

- also, before the "real" interviews, a pilot interview was conducted with each participant to explain the interview format and to give them practice in "solving and talking";
- the problems were administered to participants in different order, to mitigate possible effects of mutual influence of the problems;
- the interviews were kept short enough (45-50 min.) to ensure that fatigue was not a significant factor in the participants' performance (Miller, 1991).
- After the interview, each student was explicitly reminded not to discuss the interview with *anyone*.

Data analysis.

I was mindful of the common strategies of ensuring the data quality (Maxwell, 1996a; 1996b; Miles & Huberman, 1984; Patton, 1991). I did not resort to any one method of data collection and analysis: by combining the quantitative and the qualitative methods of research, I enhanced the reliability and objectivity of my results. To verify causal relationships and explore alternative hypotheses, I used such methods as "triangulation", the use of "rich data" and of "quasi-statistical evidence" (Maxwell, 1996b; Miles & Huberman, 1984) – they are all integral to my research design.

Overall, the validity issues played a very significant role in planning and implementing the study. I am proud to say that I seem to have exceeded the standards of validity that have long been accepted in physics education research.