

CHAPTER 3  
SUMMARY OF RESULTS AND DATA ANALYSIS

**Introduction.**

In this chapter, I present the results of the study. The readers may recall that the general purpose of the study was dual: to determine the significant factors associated with successful problem solving and to identify some teachable approaches utilized by good problem solvers. Both quantitative and qualitative methods were used to ensure the comprehensive analysis of the students' performance.

The quantitative data collected in the study were analyzed by various statistical methods to determine the predictors of success in problem solving. "Success" was measured as the number of free-response CyberTutor problems solved. The factors considered included the background information drawn from the student questionnaires, the student scores on the Mechanics Diagnostic Test (MDT) and the data pertinent to solving the free-response problems (number of hints used, number of incorrect attempts used).

The qualitative data included about two hundred written comments made by some students as they "exited" the free-response CyberTutor tasks and the transcripts of 32 think-aloud protocols of solving challenging problems. This chapter presents the narrative summary and cluster analysis of these data. The qualitative analysis provided the insights into the students' thinking as they attempted to solve challenging problems. In Chapter 4, the results of both quantitative and qualitative analysis will be discussed in conjunction.

**Background Questionnaires and CyberTutor tasks: Statistical Analysis.**

The table below explains the nature of the variables and the method of coding.

Table 1. Explanation of variables.

<b>Class</b>	<b>A group taught by one teacher; not necessarily one section at their school.</b> There were nine teacher participants and nine classes, coded, arbitrarily, from 1 to 9.
<b>Gender</b>	<b>Participant's gender.</b> Male (1) or Female (2)
<b>Math-course</b>	<b>Current math course</b> 1 – algebra or trigonometry; 2 – precalculus; 3 – Calculus AB; 4 - Calculus BC; 5- beyond Calculus BC
<b>Math-grade</b>	<b>Current math grade</b> 2 – D; 3 – C; 4 – B; 5 - A
<b>Prev#math-course</b>	<b>Last year's math course</b> Same coding as for the current math course
<b>Prev#mathgrade</b>	<b>Last year's math grade</b> Same coding as for the current math grade
<b>Physics before?</b>	<b>Prior experience with physics</b> 1 – had physics before 2- AP Physics is the first physics course
<b>Puzzles</b>	<b>Fondness for puzzles</b> 4 - solves often; 3 - solves occasionally; 2 - solves rarely; 1 – never solves
<b>Parent ed</b>	<b>Highest degree reached by either of the parents/guardians</b> 1 – high school diploma; 2 – college undergraduate; 3 – graduate/professional/doctoral
<b>Parent-pme?</b>	<b>Parents' professional background in physics/math/engineering</b> 1 – none; 2 – one parent only; 3 – both parents
<b>SAT-Math</b>	<b>Best SAT-Math score</b>
<b>MDT score</b>	<b>Score earned on Mechanics Diagnostic Test (MDT) (%)</b>
<b>MDT-R score</b>	<b>Score earned on the R-subtest of MDT (%)</b>
<b>MDT-B score</b>	<b>Score earned on the B-subtest of MDT (%)</b>
<b>Problems solved</b>	<b>Number of free-response CyberTutor problems solved (0-5)</b>
<b>Incorrect attempts</b>	<b>Number of incorrect answers submitted</b>
<b>Hints used</b>	<b>Number of hints used</b>
<b>R-Hints used</b>	<b>Number of R-hints used</b>
<b>B-hints used</b>	<b>Number of B-hints used</b>

*Note:* To achieve a sharper distinction between the influence of R- and B-questions on the number of free-response tasks solved by the participant, I considered 10 easiest questions (all of which had been classified as R-type) and 10 hardest ones (all of which had been classified as B-type). The difficulty level for the questions in these subtests was sharply different: the individual R-questions were answered correctly by 73-96% of the students, whereas for the individual B-questions these numbers stand at 15-46%.

These groups of ten questions were considered as subtests of the MDT. The student scores on these subtests were named “R-score” and “B-score”, respectively.

#### Descriptive statistics.

To get a general sense of my results, I obtained the descriptive statistics of the data. For the readers’ convenience, the categorical variables, where the numeric values were assigned arbitrarily, and the quantitative variables, for which the numeric values have significance, are organized separately: see Table 2a and Table 2b.

Table 2a. Categorical variables. Frequency and the means for the entire sample<sup>1</sup>.

	N	1	2	3	4	5	mean
Math-course	100	n/a	n/a	13	71	16	4.03
Math-grade	98	n/a	n/a	7	45	46	4.55
Prev. math course	103	n/a	77	8	16	2	2.45
Prev. math grade	100	n/a	1	6	43	50	4.60
Physics before?	103	49	54	n/a	n/a	n/a	1.52
Puzzles	103	14	42	40	7	n/a	2.39
Parent ed	103	4	15	84	n/a	n/a	2.78
Parent-pme?	103	62	28	13	n/a	n/a	1.52

Table 2b. Quantitative variables. Descriptive statistics for the entire sample.

	N	Minimum	Maximum	Mean	Std. Deviation
SAT-Math	101	560.00	800.00	748.02	44.61
MDT score	98	18.00	96.00	58.08	15.35
MDT-R score	98	30.00	100.00	84.69	14.87
MDT-B score	98	.00	100.00	33.98	21.43
Problems solved	95	.00	5.00	2.25	1.94
Incorrect attempts	95	.00	26.00	8.38	6.58
Hints used	95	.00	29.00	11.19	7.40
R-Hints used	95	.00	12.00	3.54	2.88
B-hints used	95	.00	21.00	7.66	5.07

As evident from Tables 2a and 2b, the level of completion was quite high: only eight participants failed to attempt the free-response problems, and only five did not take MDT. The background information was also provided almost without omissions, after some persistent “remote chasing” on my part. Two participants had never taken the SAT test, and three were not taking any math course that year.

What do these statistics tell us? The “average” participant was taking a BC Calculus course, doing quite well (B+/A-). He or she had had a similar grade in mathematics the year before. The participants’ mean SAT-Math score is about 750 (maximum 800), placing them, not surprisingly, in the top tier of their peers with respect to mathematical reasoning. Meanwhile, the mean SAT-Math score is close to that of the *national* sample of the students who take AP Physics C course<sup>2</sup>, thus further reaffirming the validity of the sample construction method.

In addition, the average participant enjoys an occasional puzzle – but no more than that. For about half of the participants, the AP course was also their first course in physics. The parents of an average participant are not so likely to have any background in physics-related fields (in fact, 61 of the participants answered

<sup>1</sup> The entire sample included 38 female and 65 male participants.

<sup>2</sup> Relevant statistical information can be found at

<http://www.collegeboard.com/research/abstract/0,1273,3861,00.html>

“none” to this question) – however, the level of parent education is quite high: most participants have parents who went beyond the undergraduate degree in college.

The CyberTutor tasks proved to be quite challenging for the participants: the “average” participant earned 58% on the MDT, scoring about 85% on the MDT-R-subtest and only 34% on the MDT-B-subtest (note the sharp difference here). Also, on average, only 2.24 problems, out of five, were solved, despite the hints and the ample time provided. The students did try, however: an average student used more than eight incorrect attempts and more than eleven hints in the process of solving the open response problems. B-hints were used more than twice as frequently as the R-hints were; however, there were more B-hints provided, so I did not read much into this comparison.

#### The gender factor.

Gender has long been known to be an important factor associated with success in problem solving and, more specifically, in physics and mathematics<sup>3</sup>. Being aware of that, I conducted the two-tailed t-test to compare the means for males and females. The purpose was to see if gender may, indeed, be a factor in my sample. Table 3 illustrates the results.

Table 3. Group means by gender. Asterisk (\*) notes the means that are significantly different. ( $p < 0.05$ ).

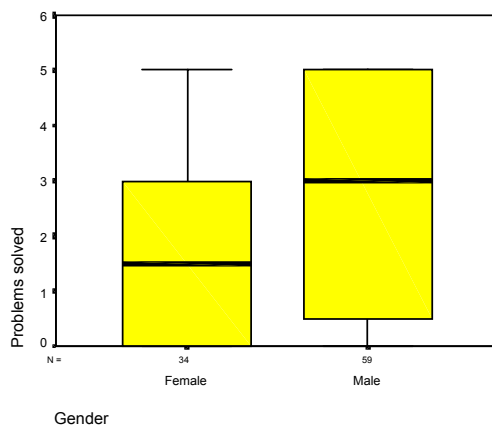
	Gender	N	Mean	Std. Deviation	Std. Error Mean	t-test sig.	Eff. size (in SD)
Math-course	Female	37	3.97	.60	.10	0.42	0.16
	Male	63	4.06	.50	.06		
Math-grade	Female	35	4.54	.50	.09	0.97	0.02
	Male	63	4.55	.54	.07		
Prev#math-course	Female	38	2.47	.79	.13	0.80	0.05
	Male	65	2.43	.84	.11		
Prev#mathgrade	Female	36	4.69	.38	.06	0.18	0.30
	Male	64	4.54	.63	.08		
Physics before?	Female	38	1.42	.50	.08	0.11	0.24
	Male	65	1.58	.50	.06		
Puzzles	Female	38	2.37	.85	.14	0.85	0.04
	Male	65	2.40	.79	.10		
Parent ed	Female	38	2.76	.54	.09	0.84	0.04
	Male	65	2.78	.48	.06		
Parent-pme?	Female	38	1.58	.76	.12	0.55	0.12
	Male	65	1.49	.69	.09		
SAT-Math	Female	37	741.35	43.15	7.09	0.26	0.23
	Male	64	751.88	45.32	5.66		
MDT score	Female	37	55.05	15.38	2.53	0.13	0.32
	Male	61	59.92	15.17	1.94		
MDT-R score*	Female	37	80.54	15.45	2.54	0.03	0.50
	Male	61	87.21	14.04	1.80		
MDT-B score	Female	37	32.70	20.09	3.30	0.65	0.10
	Male	61	34.75	22.33	2.86		
Problems solved*	Female	36	1.76	1.74	.30	0.04	0.41
	Male	59	2.52	2.00	.26		
Incorrect attempts	Female	36	9.21	6.82	1.17	0.32	0.20
	Male	59	7.90	6.46	.84		
Hints used	Female	36	13.06	8.88	1.52	0.08	0.39
	Male	59	10.12	6.23	.81		
R-Hints used	Female	36	4.18	3.47	.60	0.09	0.34
	Male	59	3.17	2.42	.32		
B-hints used	Female	36	8.88	5.94	1.02	0.09	0.37
	Male	59	6.95	4.40	.57		

<sup>3</sup> See, for instance, [www.collegeboard.com/ap/techman](http://www.collegeboard.com/ap/techman) for the information on the performance of males and females on the AP Examinations.

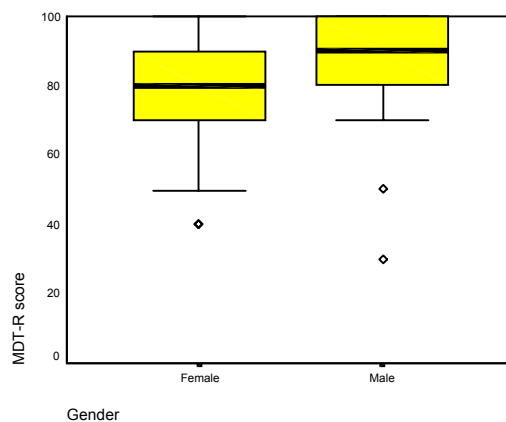
The last column in the table shows the effect size: the difference in the means expressed in terms of the mean standard deviation (SD). Consider, for instance, the SAT-Math score. The difference between the male and female mean scores is  $(751.88 - 741.35) = 10.53$ . The mean SD for the SAT-Math score is  $(43.15 + 43.31)/2 = 43.23$ . Therefore, the effect size is  $10.53/43.23 = 0.23SD$ . This is a relatively substantial effect; however, it is not significant ( $p < 0.26$ ). Using a larger sample in a future study would increase the statistical power of the t-test, rendering an effect of such size significant. In my study, however, only the effects exceeding  $0.4SD$  were significant (see Table 3).

What other observations can be made from these data? Notably, none of the background variables were significantly different for males and females; in fact, all of the background means are remarkably similar, showing that those females that make it to the advanced physics courses are, in general, not so different from males in terms of family circumstances and basics performance indicators. However, significant differences related to the CyberTutor tasks did exist: in the number of problems solved ( $p < 0.042$ ) and in the MDT-R score ( $p < 0.031$ ). On the next page, I present the box-plot graphs for the two significantly different variables.

Number of problems solved, by gender.



MDT-R score, by gender.



For the number of Problems Solved, the median value for the males was 3 problems, with the top 25% of the males solving all 5 problems. The corresponding median value for the female participants was only 1.5 problems, and very few females solved all 5 problems. For the MDT-R score, the median scores for the females and the males were 90% and 80%, respectively. Virtually no males earned less than 70% of the maximum score, whereas about 25% of the female participants did.

These comparisons reinforce the significance of the differences found by the t-test and suggest that, in addition to the whole-sample regression analysis, separate analysis for males and females be conducted.

### Regression analysis: model-building methods.

In regression analysis, three methods of model building were used, as suggested in *SPSS Base 9.0: Application Guide* (1999). These methods include:

*Forward selection*: the variables are added one by one, generally in the order of decreasing correlation with the dependent variable.

*Backward elimination*: all variables are included in the initial model and are then eliminated – generally, in the order of increasing significance.

*Stepwise method*: in this method, the previously removed variables may be reintroduced into the model; this method of “trial and error” considers relatively large numbers of combinations of variables; this analysis, involving a large number of intermediate steps, is performed by the computer.

All three methods were used to build the models for the entire sample and for the “males-only” and “females-only” sub-samples.

Note: as a preliminary step, correlation analysis of the variables was conducted for the entire sample and for the single-gender sub-samples. The results of that analysis were used to identify possible variables of interest to use in building the regression models. While *all* variables were used in the stepwise and backward-elimination models, it was comforting to see that the variables that appeared in the final regression models were the ones that were highly correlated with the outcome variable (Problems Solved) in the first place. The results of the correlation analysis are shown in Appendix G<sup>4</sup>.

### Regression analysis: the no-interaction models.

In this method of analysis, all variables described above, were used as independent variables in the linear regression models; the number of free-response CyberTutor problems solved was the dependent variable. The insignificant variables were eliminated from the model until only the predictors with  $p < 0.05$  remained. In these models, no interactions between the variables were explored.

### Entire sample: forward selection model

Table 4: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.607	.369	.338	1.57

Predictors: (Constant), Math-SAT, MDT-R score, Incorrect attempts, MDT-B score

Table 5: Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-8.407	3.488		-2.410	.018
MDT-R score	2.948E-02	.012	.233	2.395	.019
MDT-B score	2.460E-02	.009	.276	2.792	.007
Incorrect attempts	-8.945E-02	.026	-.307	-3.417	.001
Math-SAT	1.072E-02	.005	.215	2.364	.020

Dependent Variable: Problems solved

### Entire sample: stepwise method.

This method, after several attempts with varied combinations of variables, repeatedly yielded the same model as the forward selection method.

To sum up: the model indicates that 33.8% of the variance in the number of problems solved can be explained by the variance of the variables shown above, as indicated by the value of “adjusted R-square.” The MDT-B score and the number of incorrect attempts appear to be the most significant predictors ( $p < 0.01$ ). The standardized beta coefficients show the change in the dependent variable (expressed in the

<sup>4</sup> For the readers’ convenience, some of the variables that did not appear in any regression models are omitted from the correlation tables in Appendix G; however, the correlation tests were actually conducted for *all* variables.

units of standard deviation, or SD) associated with the change of the independent variable of one SD. In terms of the beta values, the “number of incorrect attempts” appears to be the most powerful predictor. The beta value for this variable is negative: the more incorrect attempts a student made, the fewer problems that student tended to solve. Other predictor variables have positive beta values: higher R-score, B-score and SAT-Math score are associated with higher numbers of problems solved.

Entire sample: backward elimination model.

Table 6: Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.581	.338	.314	1.59

Predictors: (Constant), Incorrect attempts, MDT-B score, Gender

Table 7: Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.587	.695		.844	.401
Gender	.836	.354	.213	2.362	.021
MDT-B score	3.549E-02	.008	.405	4.489	.000
Incorrect attempts	-.104	.026	-.364	-4.028	.000

Dependent Variable: Problems solved

This model, with slightly weaker predictive power, replaces MDT-B score and SAT-Math score by “gender.” The model shows that 31.4% of the variance in the number of problems solved can be explained by the variance of the variables shown above. The MDT-B score and the number of incorrect attempts appear to be the even more significant predictors in this model. ( $p < 0.0005$ ).

These two regression models for the entire sample have similar predictive power, but they both are of interest, since one of them replaces two performance variables by one demographic variable; both models will be discussed in Chapter 4.

Males only: forward selection model.

Table 8: Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.505	.255	.226	1.74

Predictors: (Constant), Incorrect attempts, MDT-B score

Table 9: Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	2.511	.540		4.645	.000
MDT-B score	2.979E-02	.011	.343	2.815	.007
Incorrect attempts	-.110	.036	-.369	-3.023	.004

Dependent Variable: Problems solved

Males only: backward elimination and stepwise method models.

These methods yielded the same model as the forward selection method, thus making me fairly confident that it was, indeed, the best possible no-interaction model.

In summary: both predictors, the B-score and the number of incorrect attempts, were highly significant ( $p < 0.01$ ). As indicated by the t-values and the beta values, higher B-scores and lower numbers of incorrect attempts were associated with higher number of problems solved. The R-square value was noticeably smaller than that for the entire sample, suggesting that the males introduce more “noise” in the results than females do. The models obtained for the females (shown below) also appear to suggest a more “predictable” problem-solving success among the females.

Females only: forward selection model<sup>5</sup>

Table 10: Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.709	.503	.452	1.24

Predictors: (Constant), Incorrect attempts, MDT-B score, Math-SAT

Table 11: Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-9.045	4.286		-2.110	.044
MDT-B score	4.183E-02	.011	.516	3.785	.001
Math-SAT	1.354E-02	.006	.317	2.364	.025
Incorrect attempts	-9.04E-02	.033	-.372	-2.747	.010

Dependent Variable: Problems solved

The model contains three variables, all performance-related. The MDT-B score is the most significant predictor; the number of incorrect attempts has the effect similar to that in the all-sample and males-only models (more incorrect attempts is associated with less successful problem solving). Math-SAT score is also significant, unlike for the male-only sub-sample.

Females only: stepwise model

Table 12: Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.706	.499	.460	1.25

Predictors: (Constant), MDT score, Puzzles

Table 13: Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.828	1.266		.654	.519
MDT-B score	3.535E-02	.012	.424	2.974	.006
Puzzles	-.867	.259	-.477	-3.343	.003

Dependent Variable: Problems solved

<sup>5</sup> One variable that was initially included in the models was Grade. This variable was a significant predictor in the original models. However, upon closer inspection, it was excluded from the models. Only five out of 38 females were 11-graders, and the rest were seniors. Of the five 11-graders, 3 came from a top-in-the-nation math and science school; these 3 participants were unusually strong. Given these circumstances, it seemed that Grade was a variable leading to a model that could not reasonably be generalized for the larger population.

In this model, MDT-B score is, again, highly significant. However, the model is different in that a background variable (self-reported fondness for puzzles) takes the place of two performance-related variables (SAT score and the number of Incorrect Attempts). The negative values of *beta* and *t* have to do with the coding system: they, actually, mean that those who liked to solve puzzles on a regular basis tended to solve *more* problems.

Females only: backward elimination model.

Table 14: Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.781	.609	.547	1.13

Predictors: (Constant), B-hints used, MDT-B score, Math-course, Math-grade

Table 15: Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-8.349	2.861		-2.918	.007
Math-course	1.064	.368	.372	2.891	.008
Math-grade	1.246	.505	.324	2.465	.021
MDT-B score	3.085E-02	.011	.369	2.828	.009
B-hints used	-.118	.038	-.398	-3.082	.005

Dependent Variable: Problems solved

Of the three models built for the female sub-sample, the backward elimination method yielded, by far, the best result: adjusted R square value was 0.547, meaning that 54.7% of the variance in the number of problems solved can be explained by the variance of the independent variables in the model. Higher math skills, as defined by the level of math course taken and by the grade in that course, are associated with a higher number of problems solved. Also, two bisociation-related factors were significant: the higher B-score and the lower number of B-hints used were associated with a higher number of problems solved.

Regression analysis: two-way interactions.

Sometimes, the independent variables have a combined effect more significant than that of the individual variables. For instance, for males, the Math-SAT score and the number of B-hints used were not significant. However, one might hypothesize that, for instance, the students who had very high SAT scores *and* used many B-hints solved the free response problems especially well. To check for such interactions, a number of new variables were computed; each new variable was a product of two original variables (two-way interactions). The new variables were combined with the original ones, and regression analysis was conducted for the entire sample, and the males-only and females-only sub-samples. In each case, all three methods (forward selection, stepwise and backward elimination) were used in an attempt to improve the models achieved with just “raw” variables. Since there were many highly-correlated variables among the newly created ones, special attention was paid to retaining only the variables with fairly high tolerance levels. After extensive research, it was discovered that the predictive power of the interactions models – as determined by the adjusted R square value – is somewhat *less* than that of the corresponding no-interactions model (see the table below).

Table 16: Comparison of the adjusted R square values for the no-interactions and interaction models

	Entire Sample	Males only	Females only
No interactions	0.338	0.226	0.547
Two-way interactions	0.320	0.213	0.506

Therefore, there appears to be no significant interactions of variables that help improve the models and, therefore, no particular gain in discussing the interaction models further.

To facilitate further discussion and to analyze the differences between the models, it is useful to display the main characteristics of these models in one table. Table 16a represents the comparative data of the models with the greatest predictive power (as measured by the R-squared value).

Table 16a: Comparative characteristics of the regression models<sup>6</sup>

	All sample (forward)	All sample (backward)	Males (forward, backward)	Females (backward) <sup>7</sup>
R <sup>2</sup> adjusted	0.338	0.338	0.226	0.547
constant	-8.4	0.587	2.511	-8.349
MDT-R	0.233			
MDT-B	0.276	0.405	0.343	0.369
Inc. attempts	-0.307	-0.364	-0.369	
Math SAT	0.215			
Gender		0.213		
Math course				0.372
Math grade				0.324
B-hints used				-0.398

As one can see, the only variable that appears in all four “successful” models is MDT-B score. The models for the entire sample have similar sets of significant predictor variables, gender in the backward model replacing MDT-R score and the Math SAT score in the forward model. The models for the male sub-sample have rather weak predictive power and contain some of the variables appearing in the “whole-sample” models. Interestingly, the most successful model for the female sub-sample contains a rather different set of variables: such variables as the students’ mathematics course and grades, as well as the number of B-hints, do not appear in the “whole-sample” and the “males-only” models.

Notably, the MDT-B score is the sole variable appearing in all four models (in fact, in all six, including the less powerful models obtained for the female sub-sample – see Tables 11, 13).

#### The “last straw” statistics.

Just like the proverbial last straw that broke the camel’s back, the last hint that helped a student solve a problem was of interest. Table 17 on the next page shows, for each of the five CyberTutor tasks, the number of times any one hint served as “the last straw” – in other words, was the last hints used by the student before submitting the correct answer. In this table, the first column is the problem CyberTutor code, the second line is the hint CyberTutor code, the third line shows the number of students who solved this problem after using this last hint, and the last column indicates the hint type (R-type or B-type; in other words, rigid-knowledge-related or bisociation-related). The empty cells in the second column indicate that no hints were used before the correct answer was submitted.

For instance, consider problem 10278: 4 students solved it after using hint 10280 as the *last* hint; 6 students solved it after using hint 10281 as the *last* hint, etc.

<sup>6</sup> Rows 3-10 show the beta values for the predictor variables that are significant in each model.

<sup>7</sup> The other two models for female participants are not presented because of their notably weaker predictive power.

Table 17: Last-hint use for the free-response CyberTutor tasks.

Note: the most helpful hint for each problem marked \*\*, the second most-helpful hint marked \*.

Problem	Last Hint	Occurrences	Type
10278 <sup>8</sup>	10280	4	r
10278	10281	6*	r
10278	10282	3	r
10278	10283	5	b
10278	10284	3	b
10278	10285	8**	b
10311	10316	4*	b
10311	10325	3	r
10311	10326	1	b
10311	10327	1	b
10311	10328	4*	b
10311	10329	9**	b
9736	9790	4	r
9736	9792	6	r
9736	9794	9*	b
9736	9796	3	b
9736	9798	4	b
9736	9800	12**	b
9834	9792	1	r
9834	9794	1	r
9834	9796	5	b
9834	9798	6*	b
9834	9800	9**	b
9862	9790	4	r
9862	9792	4	r
9862	9794	2	r
9862	9796	5*	b
9862	9798	3	b
9862	9800	11**	b

The results demonstrate that the B-type hints were far more helpful in “clinching” the solution: in each problem, the hint that was helpful most of the time was of B-type; among the second-most-helpful, 5 hints out of 6 were of B-type<sup>9</sup>. Also, for each individual problem, the total number of the last B-hints exceeds the number of the last R-hints. Consider, for instance, problem 10311: the B-hints served as “last straws” 19 times, whereas the R-hints – only 3. Altogether, there were 92 occasions when a B-hint was followed by the correct answer; there were only 38 such occasions for the R-hints.

<sup>8</sup> The problem numbers and the hint numbers correspond to their CyberTutor codes; they are shown for identifying purposes only.

<sup>9</sup> Two B-hints shared the second place in problem 10311.

Free-response task solving success: internal validity check.

In addition to the validity safeguards designed prior to the experiment (see Chapter 2, Validity Issues section), a *post factum* analysis was performed to ensure that the open-response tasks reasonably measured the level of problem solving skills. If each task was, indeed, measuring the same overall ability, one would expect that successfully solving any one problem would strongly – and positively – correlate with the overall number of problems solved<sup>10</sup>. The internal validity (reliability) of the set of open-response tasks was determined by calculating the Cronbach alpha coefficient<sup>11</sup>. The value of this coefficient varies from zero (no internal consistency) to one (perfect internal consistency). According to the accepted standards, for the teacher-made tests, the minimum acceptable value of Cronbach alpha is 0.5 (Ebel & Frisbie, 1986). In my case, that value equals to 0.86. This is an excellent level of internal consistency, exceeding the standards for the tests used for placement decisions (Ebel & Frisbie, 1986).

CyberTutor tasks: written comments.

As mentioned before, after solving each task, the students were asked to comment on the solving process – more specifically, describe their difficulties and the break-through moments. The comments were strictly voluntary; the format was unspecified. On average, between 35% and 50% of the students wrote comments for each of the five problems; overall, more than two hundred comments were collected.

The possibility of collecting the student comments, facilitated by the CyberTutor software, was not a part of the original design, and I regard the data obtained from the written comments as an interesting “bonus.” On the one hand, the written comments complemented the statistical information about the CyberTutor tasks. On the other, they complemented the interview comments in an unexpectedly fruitful way: while the interview protocols reveal the nature of the *ongoing* solving process, the written comments allowed me to explore the thoughts and feelings of the students *after* having solved (or not solved) the problem.

The comments were of uneven length: from a single word (“Arrrrrghhh!” was one of my favorites), to full-page, cleverly written, sometimes semi-relevant narratives. All comments were carefully read, and the meaningful statements within the comments were identified and tabulated. The statements were then classified according to the proposed framework, as being of R-type, B-type or G-type (general)<sup>12</sup>. Within each of those types, several subtypes were identified, as shown below. Below, I present the types and subtypes of the statements, with examples<sup>13</sup>.

R-comments: these comments refer to the relevant “standard”, or rigid knowledge of basic facts (e. g. Newton’s laws) and standard algorithms (e. g. finding the vector components)

**REP:** execution of the known “physics” algorithm:

*- the speed relative to the center of mass is the speed  $v$  minus the speed of the center of mass,  $mv/(m+M)$ . The acceleration toward the center of the circle (i.e., the center of rotation, i.e., the center of mass) then was the relative velocity (of mass  $m$ ) squared divided by  $r$ , where  $r$  is the distance the center of mass is from mass  $m$ , that is,  $ML/(M+m)$ . Since the force of the string is the only force on the ball, the tension is mass times its acceleration.*

**RDP:** difficulties with the known “physics” algorithm.

<sup>10</sup> I did not conduct a similar check for the MDT multiple-choice questions since they were explicitly designed to test primarily the understanding of *different* physics concepts.

<sup>11</sup> For dichotomous data such as mine, the Cronbach alpha coefficient is equivalent to the Kuder-Richardson 20 coefficient (SPSS Base 9.0: *Application Guide*, 1999).

<sup>12</sup> Such classification proved quite convenient, as I was able to highlight the comments on the computer screen in Red, Blue and Green...

<sup>13</sup> Aside from the spelling errors, the comments are unedited – some of them are quite delightful to read, I think.

- I knew the ball of mass  $M$  -was- going to accelerate, but decided that I must be delirious from too much cold medication or something and just did it as I would a problem involving rotation around a fixed center. Then I realized my error, got a tasty snack, and set up some equations
- I picked the wrong axis at first, but I figured out that I was making it harder on myself and did it again.
- I am not sure what happened, but most likely every time I attempted to find all frictions and forces in the  $x$  direction I got a sign wrong somewhere and screwed it up
- I couldn't decide if the frictional forces between the blocks canceled out or not.
- The thing I couldn't decide right away was whether to count the mass of the 5kg block twice, because it moves twice as fast relative [to] the surface it is grinding against as the 8 kg block does against its surface.

**REM:** execution of the mathematical algorithm.

- I found that the velocity of the block was equal to  $\sqrt{48/13}$ . I multiplied that by four, and now—here's the flash of brilliance (I think)-- realize that I should have divided it by four (duh)
- I tried plugging this velocity into Newton's second law, and got the number 48 for the total force

**RDM:** difficulties with the mathematical procedure:

- I couldn't get the equations to cancel things out so that I could solve for a certain variable.
- I have all five variables and all five equations, but I just can't figure out the algebra
- I have to draw my triangles very carefully and think very carefully and re-check everything six times, and I hate working like that.

**B-comments:** these comments are related to *bisocation*, or one's ability to see the relevance of seemingly unrelated concepts to a problem situation or to *notice the trick* that helps solve the problem.

**BD:** difficulties related to bisociation:

- Hard to visualize
- I wasn't sure how to apply the concepts.
- I did not know how to set the problem up.
- I have never dealt with two mass rotating around a center of mass like this before
- I was just kind of dazzled by the diagram, and stopped thinking when I saw the hints.
- I was very worried and spent some time trying to figure out where in the problem concept I had gone so horribly wrong.
- I am in shock. I have no idea.

**BS<sup>14</sup>:** overcoming difficulties by “noticing the trick.”

- important ideas that I used were thinking about the blocks together as a system.
- once I realized that the wedge and the block can be treated as a single object, their acceleration is found easily
- the only thing that made it nontrivial was the fact that there was a lot of friction to take into account, and the normal force on the surface under block  $M$  was  $g$  times the mass of both the blocks, while the normal force on block  $M$  itself was  $g$  times the mass of block  $m$ .
- when I realized that I would have to split up the force of friction into into-the-ramp and parallel-to-the ramp directions, I realized that there HAD to be an easier way.

**BH:** overcoming difficulties with the help of hints.

- I wasn't sure that the accelerations of the two blocks would be the same until I opened that hint.
- The only comment that helped was the reminder that the frictions were working in opposite directions.

---

<sup>14</sup> S is for solving.

*- The only hint that I found helpful, and that really made the answer click in my head was number vi. I didn't realize that the centripetal acceleration was in relation to the speed of the moving ball around the center of mass.*

**G-comments:** these comments are of more general nature and related to the problem and the solver in less specific terms.

**GP:** overall quality of the problem:

**GPG** (“good” problem):

- I was impressed with the problem because it required modification of the normal force of the block because of the downward acceleration of the wedge. This was a clever question, and it is probably my favorite one so far.*
- This problem turned out not to be very inherently difficult, but the way it is set up makes it seem more difficult than it really is.*
- I thought the problem itself was pretty unique and was a good problem to tackle.*
- I thought the problem was pretty solid with no ambiguity.*
- I learned a great deal from this problem.*

**GPB** (“bad” problem):

- There was one major flaw I found in this problem. The problem was a little bit deceptive because the picture shows that the two balls have different radii.*
- It was just a bad problem.*
- The wording was confusing*
- This is one of those problems that don't let me sleep at night*

**GH:** overall quality of the hints:

**GHG** (“good” hints):

- hints actually did help... sometimes they were obvious though.*
- the hints were a little helpful*
- I, personally, would have only found hints iii through vi helpful*
- The last hint I found to be incredibly helpful.*

**GHB** (“bad” hints):

- The three hints I opened told me everything I knew*
- Once more, the hints themselves were really blunt and not very helpful.*
- I wish that the hints would tell more.*
- I found myself extremely frustrated by the fact that every time I opened a hint, I was given information I already knew and understood*

**GG:** general feelings about the solving process and/or self<sup>15</sup>.

- I'm disappointed that I didn't solve this problem sooner. I'm just glad that I was able to figure it out in the end before it was too late*
- I ended up finding  $V_p^2$  to be 12.8, which is a nice reassuring roundish number. I like those.*
- I. Hate. Ramps.*
- I decided that if there wasn't an easier way, I'd just give up and console myself with a Krispy Kreme donut.*
- I've found that I only have about two creative approaches to a given problem, and if those both fail, I'm out of the running*
- Problems like these make me want to run a red-hot cheese grater over the small of my back until I hit the parmesan.*
- If this were a test, I just bombed it! I was unprepared for the toughness of the problems.*
- I was mad beat.*
- I was rather intimidated by the problem. I had no idea that this was going to be so complicated.*

---

<sup>15</sup> These are the most fun to read, so there are more examples here.