CHAPTER 4
METHODS OF VERIFICATION

This thesis describes the development of problem-solving skills of two matched cohorts of students in two introductory college physics courses. The students in one cohort (EPS) were taught an explicit problem solving strategy while the students in the other cohort (TRD) were not. This description of the development of problem-solving skill was not a straight-forward endeavor. Not only were the results of this interpretative study exposed to many validity threats, but the details about the students within the cohorts were also important. The goal of this chapter is to use data from many different sources to add validity to the results presented in the next chapter by addressing known (or predicted) threats to validity. This technique is known as triangulation and is common in most interpretive studies.

The first threat to the validity of this study involved the four problem-solving skill measurements. Perhaps they do not measure what they were designed to measure or maybe using one skill would have been sufficient. The second threat to validity involved the use of the problem difficulty ranks to adjust the problem-solving skill development results. Just because they had face-validity does not guarantee that the difficulty ranks were actually valid. The third major threat involved the selection of the students to the cohorts. If these students were not representative of their classmates then there would be no generalizability, even locally, of the results. The final threat that will be examined involved the use of the Minnesota Problem-solving Strategy by the EPS students. If the students didn't use the strategy, the interesting difference between the EPS and TRD cohorts would vanish. This chapter will examine the study's validity.

Coding Rubric Verifications

The open-ended problem solutions generated by the students were interpreted with the help of a problem-solving skill coding rubric. Four problem solving skills were selected to be analyzed. For review, the four skills are General Approach, which was a measure of the student's conception of which physics principles to apply; Specific Application of the Physics, which was a measure of how well the students did what they thought they needed to do; Logical Progression, which was a measure of the student's planfullness; and Appropriate Mathematics, which measured the students use of math in a physics setting. There were three principle threats to validity involving the problem-solving skill measurements made using the coding rubric. The first threat was the consistency of the codes in being applied. The second threat involved the independence of the four skills. The third threat involved possible checks that the General Approach dimension was measuring something sensible. These three validity checks comprise the next three sections of this dissertation.

Intra-rater Reliability

A very straightforward, albeit necessary, check of the validity of the codes is to see if the codes can be consistently applied by the same person. In education research this is called intra-rater reliability. For this study 30 students had their solutions to problem T1-Q3 coded and after a few weeks had pasted, the same solutions were coded again. By comparing the two attempts at coding, the intra-rater reliabilities can be computed. Since the codes are ranks, the Spearman Rank Correlation will be used. Generally, correlation coefficients above $r_s = 0.8$ are considered adequate. For the General Approach skill, the correlation coefficients was $r_s = 0.90$. For the Specific Application of the Physics skill, the correlation coefficients was $r_s = 0.91$. For the Logical Progression skill, the

correlation coefficients was $r_s = 0.82$. For the Appropriate Mathematics skill, the correlation coefficients was $r_s = 0.82$. Therefore using the codes are reliable and consistent.

Coding Rubric Correlations

The problem-solving literature reviewed in Chapter Two of this study provided some theoretical support for the four problem solving skills examined and coded for each student's solution. The same literature also established that problem solving is a complex skill. It would be surprising if the four skills selected for this study were independent dimensions of a total problem-solving skill. Instead, the skills should be interrelated to each other in some manner. The validity threat was concerned with the level of this relationship. If all the skills were correlated then either too many measurements were made or the wrong measurements were made.

To check any possible dependence between the problem-solving skills, a cohort-level correlation matrix was created. The average scores (unadjusted for difficulty) received by each cohort on each dimension were correlated with each other. Since there were six correlation tests, the significance power was reduced to □=0.05/6 =0.0083. Tables 4.1 and 4.2 show the correlation matrices.

Table 4.1

Spearman Rank Correlation for EPS Cohort's Problem-solving Skills Scores.

	GA	SAP	LP	AM
General Approach	1.000			
Specific Application of Physics	0.585**	1.000		
Logical Progression	0.640**	0.766**	1.000	
Appropriate Mathematics	0.276	0.481*	0.707**	1.000

Table 4.2

Spearman Rank Correlation for TRD Cohort's Problem-solving Skills Scores.

	GA	SAP	LP	AM
General Approach	1.000			
Specific Application of Physics	0.273	1.000		
Logical Progression	0.201	0.518*	1.000	
Appropriate Mathematics	0.115	0.369	0.679**	1.000

Note. $N = 27$, $*p < 0.05$, $**p < 0.0083$

For the EPS cohort, the first significant correlation was between the General Approach score and the Specific Application of Physics score. This was re-assuring from a pedagogical perspective. A teacher would hope that the student who had chosen an inappropriate mental space would subsequently falter in applying physics in the wrong mental space. Additionally, the more incorrect the initial space was, the more difficult it should be to correctly apply the physics; and vise versa. It was worrisome that this correlation was not significant for the TRD cohort, but this lack of parity did demonstrate the validity of measuring both skills.

The next three correlation coefficients between problem-solving skills in the EPS cohort seemed to suggest the skills were not independent. The Logical Progression scores

correlated with the other three skills. However, drawing the conclusion of dependence was premature. When the correlation coefficients between problem-solving skills were computed for the more traditionally taught TRD cohort, there was only one significant correlation. This was the one between Logical Progression and Appropriate Mathematics. This result again reinforced the validity of measuring all four skills.

The different between the correlation matrices for the two cohorts was interesting as more than a validity confirmation. It highlighted a difference between the two courses. The students in the EPS cohort were taught to write their solutions as a coherent argument since the solutions would be graded by the instructor for overall planfulness. This was clearly related to the Logical Progression problem-solving skill. It should not be surprising the four problem-solving skills hung together with Logical Progression. For the TRD cohort, and as is traditional in most introductory physics course, the final answer tended to be more important that the process of getting the solution. The single correlation of Logical Progression and Appropriate Mathematics highlights the importance of the answer in the TRD course. There was a clear difference in the instruction and these correlation coefficients show that instructional differences can impact how students solve problems. This result will be revisited in later analyses.

Multiple Choice Questions

The next validity threat was the determination of the student's physics knowledge from their solutions. Several researchers have shown that the students can solve problems without knowing the physics (Halloun & Hestenes, 1985; McDermott, 1984). This conclusion has lead to the development of independent tests of student's physics knowledge, most notably the Force Concept Inventory (FCI). The FCI and the In-house Concept Tests were used in this study as independent measures of student's knowledge. In addition, the General Approach problem-solving skill measured what physics the students wanted to use to solve the problem. If these multiple choice tests did not correlate with the student's General Approach scores, then there would be concerns over the validity of the General Approach scores. This section of the thesis examined the relationship between the multiple-choice tests and the General Approach problem-solving skill

In general both cohorts did well on the multiple-choice tests. Table 4.3 displays the average percent correct scored by each cohort on each of the tests. The first two set of data points represent the pretest and posttest of the FCI. The last two points are the two in-house tests given on the second- and third-term final exams. There were no pre-tests for these two in-house tests. These tests are in Appendix E.

Table 4.3

Average Percent Correct on Multiple choice Tests for Both Cohorts

	EPS cohort			TRD cohort			
Test	M	SE	N	M	SE	N	t-value
FCI pre	48	3	24	50	4	24	1.12
FCI post	83	2	24	73	3	24	17.7*
T2 m/c	71	3	24	64	3	24	4.39*
T3 m/c	64	3	24	52	4	24	10.5*

Notes: * $p<0.05$

The only non-significant difference in Table 4.3 is the FCI pre-test, which was not surprising since the cohorts were matched using the FCI pretest. For the remainder of the tests, the EPS cohort did significantly better, even on the third-term questions (T3 m/c) which had more short-answer mathematical questions than either the FCI or the second-term multiple choice questions (T2 m/c). It was apparent from this table that the students in the EPS cohort did better on the multiple-choice tests and may have a better conceptual grasp on the physics covered by these tests.

It was clear from these results that the students had learned in their courses. What remained unclear was the relationship between these results and the General Approach problem-solving skill. The General Approach score should be correlated with the FCI score. Table 4.4 shows the correlation between the relevant final exam and multiple choice posttests.

Table 4.4

Correlation between General Approach Score on Final Exams and Multiple Choice Test Performance for Both Cohorts

	EPS cohort		TRD cohort		
Comparison	R	P	R	P	N
FCI Post, GA T1-F	0.239	0.265	0.410	0.046	24
T2 m/c, GA T2-F	0.529	0.007	0.414	0.044	24
T3 m/c, GA T3-F	0.696	<0.0001	0.804	<0.0001	24

Notes: P from Fisher's r to z

The six correlation measures show that in all but one case a correlation exists. The only non-significant correlation had an easy explanation. The students in the EPS cohort all did very well on the FCI posttest. The lowest score was 60% correct with one student getting 97% correct. With such a small variation in scores, the correlation was naturally low. This was as predicted. It would appear that the General Approach scores of the final exams do correlate with the students' performance on the multi-choice test. This result added validity to the General Approach measurements and ends the section demonstrating the overall validity of the coding rubric.

Problem Difficulty

The variable difficulty of exam problems is one of the confounding variables in any study looking at development of problem-solving ability by interpreting student answers to exam problems. Easy problems would tend to elevate problem-solving ability, while difficult problems would mask problem-solving ability. To account for this effect, this study used the difficulty rank discussed in Chapter Three to adjust the problem-solving skill score assigned to each solution based on the coding rubric. Multiplying the

rubric score on each problem by the problem's difficulty rank and then dividing by five did this adjustment. The score was divided by five because five was near the average difficulty rank for each section which would allow the adjusted scores to be in the region of the scores represented by the problem-solving skill coding rubric.

Because the coding rubric's scores were adjusted by the problem's difficulty rank a validity threat was introduced. If the difficulty ranks were not valid, then the adjustment would blur the results instead of clarifying them. It was also not clear if the simple linear adjustment was valid. The validity of using the difficulty ranks needed to be established.

Each of the twenty-one traits that make a context-rich problem difficult did have face-validity and they were supported by several research studies. In addition, it was possible using data from this study to statistically verify the validity of the difficulty traits by examining the overall difficulty rank. The difficulty ranks are presented in Table 4.5. Notice that the difficulty ranks were different for each cohort. This reflected the different instructional experience had by the two cohorts. Sometimes this experience was permission to use a different set of equations. Or this different experience manifested itself as a difference in the timing rank adjustment used to reflect the familiarity of the students with the concepts. The timing rank adjustment was discussed in the previous chapter.

Table 4.5

Problem Difficulty Ranks and Percent of Cohort with the Solution Correct

	Difficulty rank		Percent correct	
Problem	TRD	EPS	TRD	EPS
T1-Q3*	4.0	4.0	12.0	20.0
T1-Q4	2.0	3.0	36.0	22.1
T1-F1	1.5	1.5	56.0	72.0
T1-F2	4.0	4.0	40.0	36.0
T1-F3	2.5	3.5	32.0	28.0
T1-F4	4.5	4.5	28.0	44.0
T1-F5	5.0	5.0	12.0	8.0
T1-F6	6.0	6.0	12.0	36.0
T2-Q1	5.0	5.0	16.0	16.0
T2-Q2	4.0	4.0	32.0	20.0
T2-Q3	5.5	6.5	9.1	12.2
T2-F1	4.5	5.5	19.2	14.1
T2-F2	5.5	5.5	20.0	52.0
T2-F3	5.0	5.0	32.0	16.0
T2-F4	5.0	5.0	20.0	19.2
T2-F5	6.0	6.5	4.0	4.0
T2-F6	5.0	5.0	16.0	12.0
T3-Q1	6.5	6.5	4.0	4.0
T3-Q2	5.5	6.5	32.0	4.0
T3-Q3	7.0	8.0	4.0	0.0
T3-F1	7.0	7.0	4.0	0.0
T3-F2	3.5	4.5	36.0	12.0
T3-F3E	5.5	5.5	16.0	0.0
T3-F3B	3.5	4.5	48.0	20.0
T3-F4	5.0	5.0	16.0	12.0
T3-F5	8.0	8.0	8.0	4.0
T3-F6	6.0	6.0	8.0	12.0

Note: T1, T2, and T3 refer to the academic term, either first quarter, second quarter or third quarter. Q stands for quiz. F stands for final exam.

These ranks were then compared to the percentage of the students in each cohort who got each problem completely correct. Completely correct meant the students must have had a perfect solution. No partial-credit, or even silly math, errors allowed. The percentages are also shown in Table 4.5. The results of these comparisons are shown in Figures 4.1 and 4.2. Since the difficulty ranks are interval data and not continuous, it is necessary to use a non-parametric test to determine the correlation. In particular, Spearman's Rank Correlation was used. For the cohort taught an explicit problem-solving strategy (EPS cohort), the correlation between the percent correct and the difficulty ranks was r_s=-0.72. For the more traditionally taught cohort (TRD cohort), the correlation between the percent correct and the difficulty ranks was r_s=-0.85.

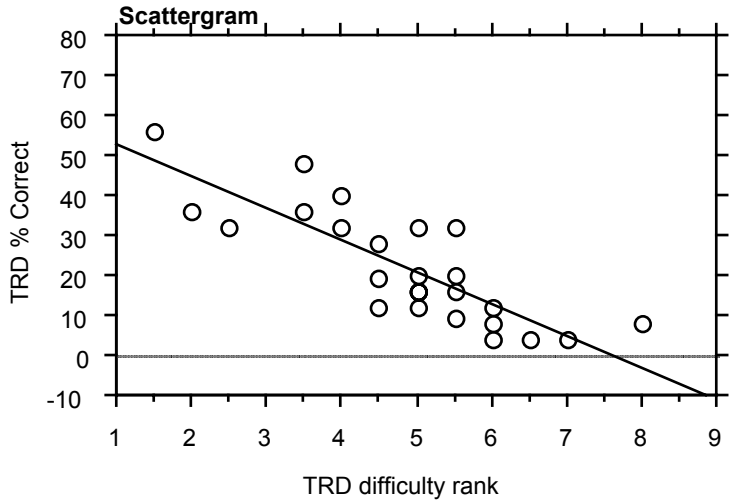

Figure 4.1. Difficulty ranks of the problems versus percent of cohort that got the solution absolutely correct for the more traditionally taught (TRD) cohort.

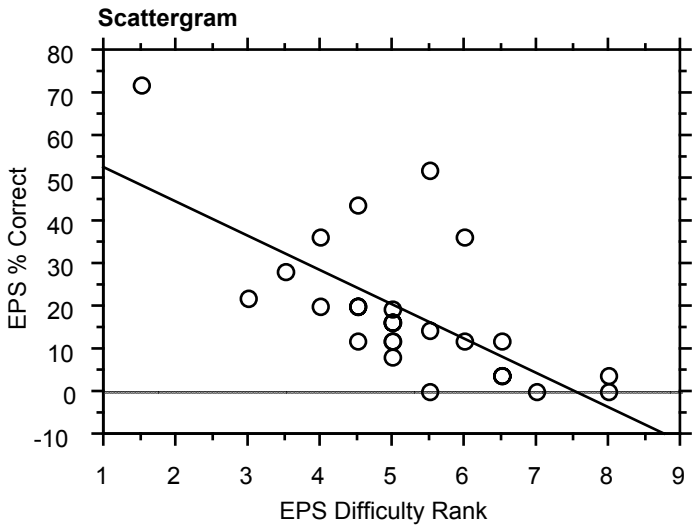

Figure 4.2. Difficulty ranks of the problems versus percent of cohort that got the solution absolutely correct for the explicit problem-solving instruction (EPS) cohort.

These are very good coefficients demonstrating that the difficulty ranks are good linear predictors of student performance on the problems in both cohorts. Since the difficulty ranks are valid, then the adjustment of the codes by the difficulty ranks is allowed. The validity of using the difficulty ranks has been established.

Cohort Representativeness

The third major threat involved the selection of the students to the cohorts. The purpose of the matched cohort was to create and follow two evenly matched teams of students through each of the two instructor's courses. In effect the intent was to observe what would happen to the same students if they could have experienced different introductory physics courses with different emphases on problem-solving instruction. There was an implicit assumption that the two cohorts would be representative of their fellow students in each course, or at least those students who remained with the same instructor throughout the academic year. Without meeting this assumption, issues of

sampling-bias could arise. The two cohorts will be compared to their classmates with their initial demographic information (the matching parameters) and overall course grade assigned to each student. Both data sets provide verification of the representativeness assumption and will be discussed in the following sections.

When these comparisons began, it was observed that one of the students in the EPS cohort had received a "D" in the second-term class. This grade was unacceptably low since it represents very little effort by the student and was not representative of the class. This student was removed from the EPS cohort as was his match partner in the TRD cohort.

Populations of Students

Each class had four different populations of students within them. The first population was the 24 students who comprised the *cohort*. The next population was the students who had stayed with the same professor for the full academic year, but were not a part of the cohort. This population was called the *remainder* since the cohort was drawn from them and the remainder population was what was left. The next population was the *survivors*. These students had completed the introductory physics course in one contiguous year, but with different professors. The final population of students was dubbed *other* since these students did not complete the course in one year. The students in the *other* population had either dropped-out of the sequence or dropped-into it. Table 4.6 shows the number of students in each population for the three academic terms (quarters). Appendix A contains a more complete breakdown.

Table 4.6

Number of Students in Each Population of Students

	EPS Course			TRD Course		
Population	T1	T2	T3	T1	T2	T3
Cohort	24	24	24	24	24	24
Remainder	52	52	52	16	16	16
Survivor	67	84	95	54	78	72
Other	79	69	54	54	47	37

Table 4.6 clearly shows that the EPS class was the larger of the two classes. However since students self-select for which section they will enroll, it is difficult to make many inferences from the relative sizes of each population. Additionally, the *other* population was ill-defined and was not a population of students the cohort was trying to represent. Therefore, this population will be ignored. Furthermore, the survivor population can be sub-divided into two groups; those students who had the same instructors for two of the three terms and those students who only visited the instructor's class for one academic term. The number of students who contiguously completed the three academic-term sequence are shown in Table 4.7.

Table 4.7

Number of Survivor Students by the Number of Terms Spent With an Instructor

	EPS Course			TRD Course		
Population	1251	1252	1253	1251	1252	1253
2-Terms	38	48	46	24	46	36
1-Term	29	36	49	30	23	36

Analysis of the Matching Parameters

The students from both cohorts were matched with each other on several pre-test measures with the goal of following these "identical" students through different problem-solving instruction. It was assumed that the cohort students were representative of their peers on these matching parameters. This section of the chapter examines this assumption through a series of ANOVA calculations where the population was the independent variable.

Review of Matching Parameters

The Force Concept Inventory (FCI) was one of the principle matching parameters as it carried three times the base rate discussed in Chapter 3. Using the FCI as a placement test was encouraged by the FCI authors (Hestenes, Wells, & Swackhamer, 1992), although disputed by others (Huffman & Heller, 1995a). In spite of this controversy, all of the problem-solving literature reviewed in Chapter 2 highlights the importance of conceptual physics knowledge. Pragmatically, the FCI offered the best choice to measure initial physics knowledge.

The Maryland Physics Expectation Survey (MPEX) created by Redish, Saul, and Steinberg (1998) was used as a gauge of the students' general attitude of the course. The

MPEX was double weighted as a matching factor. The reason the MPEX was used was to prevent matching a disillusioned student with an enthusiastic student and vise versa.

The students' background preparation was also used as a matching parameter encompassing both their previous math courses and previous physics classes. In addition, the number of hours the student was employed, their major, and their sex were also matching parameters. The problem-solving skills used as matching parameters are not included in this analysis since data only exists for the cohort and remainder students.

Matching Parameter Comparisons

To determine if the cohorts were representative of their classmates based on their matching parameters, seven ANOVA calculations were run for each class. The ANOVA compared the three interesting populations (cohort, remainder, and combined survivors) to each matching parameter. Table 4.8 is the results for the EPS class. Table 4.9 is for the TRD class.

From Table 4.8, there was only significant difference between the populations. For the student's major the EPS cohort tended to be more technical (engineering or physics) than either of the other two populations. Since the EPS cohorts major was determined by the TRD cohort this result is not detrimental to the representativeness of the EPS cohort.

Table 4.8

Analysis of Variance for matching parameters by population EPS class.

EPS	FCI		MPEX		Math Prep		Phys Prep	
	df†	F	df†	F	df†	F	df†	F
Population	2	0.40	2	0.35	2	0.59	2	2.33
Residual	120		119		120		120	

	Work		Major		Sex	
	df†	F	df†	F	df†	F
Population	2	0.84	2	4.18*	2	0.12
Residual	92		139		139	

Notes: † changes in df are due to missing or incomplete data

* $p < 0.05$

Table 4.9

Analysis of Variance for matching parameters by population TRD class.

TRD	FCI		MPEX		Math Prep		Phys Prep	
	df†	F	df†	F	df†	F	df†	F
Population	2	0.91	2	1.46	2	0.23	2	3.66*
Residual	86		78		86		86	

	Work		Major		Sex	
	df†	F	df†	F	df†	F
Population	2	0.25	2	2.10	2	3.14*
Residual	42		91		91	

Notes: † changes in df is due to missing or incomplete data

* $p < 0.05$

From Table 4.9 there are two matching parameters which are significantly different between populations; sex and physics prep (background). Both differences are statistical aberrations. In the case of sex, this aberration is from the small remainder population having an equal number of men and women, instead of the usual 80%-20% split. Since this difference is in the remainder population and the survivor population has the expected sex ratio, this difference does not change the representativeness of the TRD cohort. In the case of physics prep, all of the students in the TRD cohort had only high school physics. The lack of variance in this measure raises doubts about the computed results.

The ANOVA results presented in Tables 4.8 and 4.9 demonstrate that the EPS cohort and TRD cohorts are representative of the respective classmates. Using an ANOVA did not violate the spirit of the case-study methodology since the statistics were only used within the same class and not between cases. The next sections examines if the cohorts are representative based on the grades they received.

Grades

As was discussed in Chapter Three, overall grades were assigned to the students for their work completed in the classes. Both instructors published the same grade divisions to assign the students' overall grades. These are shown again in Table 4.10. However, closer inspection of the grade books showed that these were not rigidly followed. The actual grade divisions are shown in Table 4.11. The differences seen in Table 4.11 between each class and from the published guidelines were not done with malice. Rather the intent was to be certain all the lecture sections gave proportionally about the same grades.

Table 4.10

Published Percent of Total Possible Points Necessary to Receive Each Grade.

	Course		
Grade	T1	T2	T3
A	80-100 %	83-100 %	81-100 %
B	70-79 %	70-82 %	70-80 %
C	50-69 %	50-69 %	50-69 %
D	40-49 %	40-49 %	40-49 %
F	0-39 %	0-39 %	0-39 %

Table 4.11

Actual Percent of Total Possible Points Necessary to Receive Each Grade.

	EPS Course			TRD Course		
Grade	T1	T2	T3	T1	T2	T3
A	79.5	79.6	78.5	80	81.3	78.4
B	69.5	69.5	68.5	70	69.9	64.8
C	50	49.5	50	50	49.3	40
D	40	41	47	40	39	38

There were many differences in the grading between the two courses. Since the teaching assistants (TAs) were different people for each lecturer, the grading done by each TA was different. More importantly, each course emphasized different goals in the grading. As was already mentioned, the EPS class emphasized clear presentation of the material. In the EPS class, the explicit problem-solving strategy was reinforced by grading for a logical presentation. In the TRD class, the emphasis was placed on getting toward the right answer. However, the overall course grades still provided a

measurement of the quality of the students, provided the similarities and differences in grading were properly accounted.

Within-class grade cohort comparisons

The use of the course grades was a straight-forward method to check the representativeness of the cohort to the remainder of each class. The use of course grades was valid since the differences in the two instructors was irrelevant to within-class comparisons. What was assumed was that the grading was internally consistent within each class.

In order to check the representativeness of the cohorts to their class, the percent of total possible points was plotted for each of the three academic terms for the cohort, remainder and both survivor populations (two terms the same and one term the same). Furthermore, t-tests were computed for each of the six population pairings and the power reduced by a factor of six since these are post-hoc t-tests.

The analysis begins with the EPS cohort. Figure 4.3 shows how the three populations in the EPS class compared on final grade received in each course by term. Table 4.10 reports the numerical data.

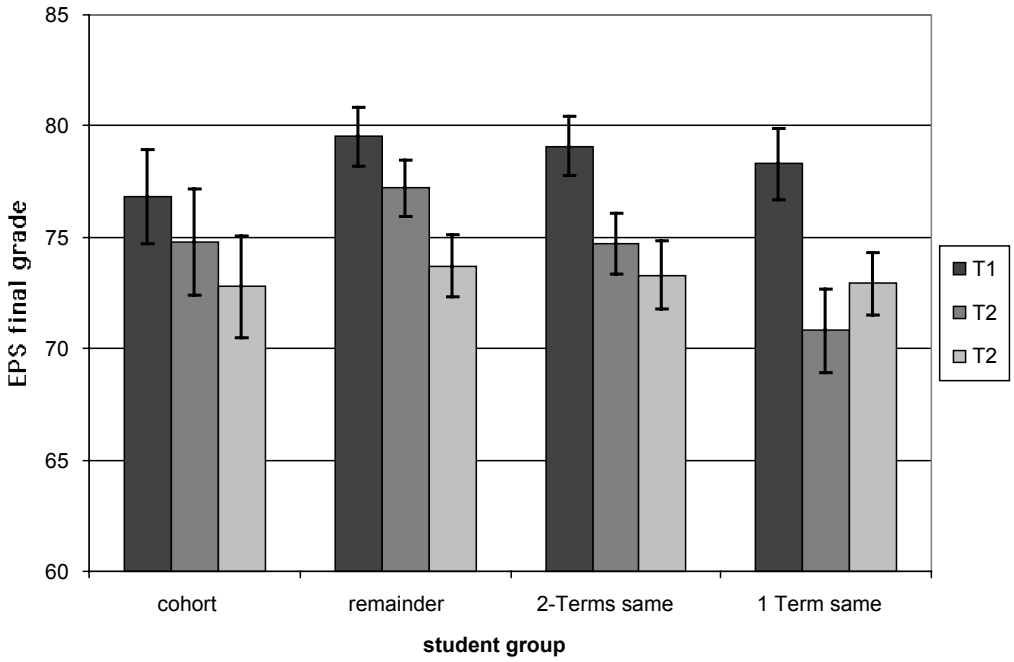

Figure 4.3. Average percent of possible grade for the three academic terms in the EPS class. The three populations of surviving students are shown.

From Figure 4.3 and Table 4.10 it was clear that the cohort population was very representative of the survivor populations in the third term and a slight under-representation of the other populations in the first term (T1). This was not a statistically significant difference. The results of the t-tests are shown in Table 4.12. The low average score seen in the One-Term population during the second academic quarter (T2) might be the result of those students failing to pick-up the problem-solving strategy. Overall, the EPS cohort appeared to be representative of those students who completed the course.

Table 4.10
Data for Figure 4.3 showing standard errors and counts for EPS class.

EPS			T1	
	cohort	remainder	2-Terms same	1 Term same
N	24	52	38	29
M	76.85	79.53	79.10	78.32
SE	2.12	1.34	1.33	1.60
			T2	
	cohort	remainder	2-Terms same	1 Term same
N	24	52	48	36
M	74.78	77.21	74.71	70.80
SE	2.392	1.24	1.37	1.88
			T3	
	cohort	remainder	2-Terms same	1 Term same
N	24	52	46	49
M	72.79	73.71	73.30	72.91
SE	2.271	1.39	1.54	1.39

Figure 4.4 shows how the four populations in the TRD class compared by term. Table 4.11 shows the numerical data. From Figure 4.4 it was seen that the cohort population remained consistently higher than all the populations for all three classes. This difference was significant between the cohort and the population of students who had the TRD instructor for two terms during the first term class. Why during T1 having two terms with the same instructor population performed so much lower than expected was not clear. The important observation from Figure 4.4 was that, although not generally significantly different, the TRD cohort might be an overestimate of the class they are representing, at least in grade performance.

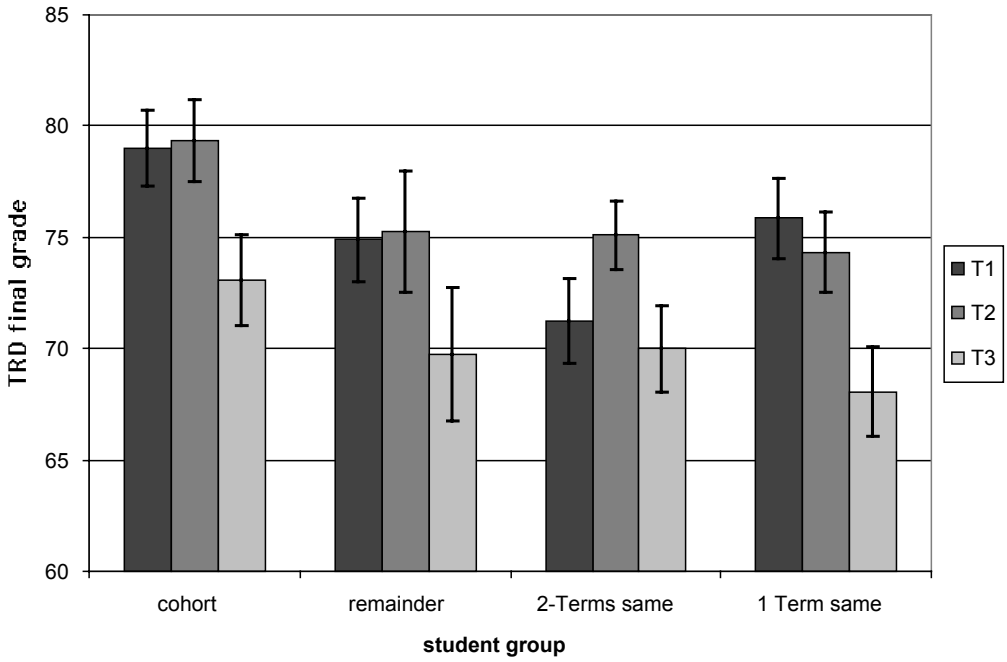

Figure 4.4. Average percent of possible grade for the three academic terms for the TRD class. The three populations of surviving students are shown.

Table 4.11
Data for Figure 4.4 showing standard errors and counts for TRD class.

TRD			T1	
	cohort	remainder	2-Terms same	1 Term same
N	24	16	24	30
M	79.01	74.90	71.24	75.85
SE	1.68	1.86	1.91	1.81
			T2	
	cohort	remainder	2-Terms same	1 Term same
N	24	16	46	32
M	79.35	75.28	75.10	74.32
SE	1.85	2.72	1.54	1.81
			T3	
	cohort	remainder	2-Terms same	1 Term same
N	24	16	36	36
M	73.08	69.74	69.98	68.06
SE	2.03	2.97	1.96	2.00

Table 4.12
Results of t-tests comparing the four populations: t-values shown.

	EPS Course			TRD Course		
Comparison	T1	T2	T3	T1	T2	T3
Cohort vs. Remainder	1.071	0.909	0.346	1.637	1.244	0.928
Cohort vs. 2-Terms Same	0.909	0.025	0.184	3.218**	1.812	1.100
Cohort vs. 1-Term Same	0.557	1.329	0.045	1.196	1.998	1.767
Remainder vs. 2-Terms Same	0.229	1.384	0.200	1.390	0.056	0.068
Remainder vs. 1-Term Same	0.586	3.032**	0.407	0.367	0.292	0.468
2-Terms Same vs. 1-Term Same	0.379	1.729	0.186	1.794	2.149*	0.687

Note. *$p < 0.05$, **$p < 0.008$

Grade Tiers

Looking at the total final grades received by the students in the cohort gave one indication that the TRD cohort may not be representative of the other populations. The letter grade received by each student gave another more graphic realization that the TRD cohort was populated by better-than-average students. This is seen with the help of a Chi-square analysis where an even distribution of eight students per grade was assumed. The number of students in each grade is shown in Table 4.13 along with a Chi-square analysis.

Table 4.13

Number of Students in Each Tier of Grade

	EPS Course			TRD Course		
Grade	T1	T2	T3	T1	T2	T3
A	11	9	7	14	10	8
B	5	7	8	6	12	12
C	8	8	9	4	2	4
χ^2	0.325	0.882	0.882	0.030*	0.030*	0.135

Note. *$p < 0.05$,

Table 4.13 shows the TRD cohort was not a balanced sample of A, B, and C students for the first two terms. Rather, the TRD was top heavy until the last term. In fact only about a third of the C students re-enrolled with the TRD instructor after the first and second terms. This is compared to a 50% re-enrollment rate for C students in the EPS course. There were simply no C students in the TRD course to be eligible for the cohort.

Another question stemming from Table 4.13 involved the stability of the students within the tiers. In the EPS cohort, 10 students stayed within their grade tier for all three terms. Of the 14 students who switched, three students improved and eight declined. The others three in the EPS cohort stayed about the same on average. In the TRD cohort, 13 students stayed within their tier. Of the 11 students who switched, two improved while seven declined. The two remaining students in the TRD cohort stayed about the same. This quick counting showed there were no hidden transitions in Table 4.13. Both cohorts had about the same improvement and decline in grades during the year.

An important use of the total grades of the students is shown in Appendix D. Here the student's total grades provide a source of triangulation data for the problem-solving skill coding rubric. This auxiliary validity check provided evidence that the rubric scored solutions in a sensible way. The analysis is relegated to an appendix since the analysis was used strictly for triangulation and its presentation in this main body of this thesis might be confusing. The value of this analysis is the conclusion that the coding rubric is loosely related to total grades.

Summary of cohort grade representativeness

As far as grade representativeness is concerned, the EPS cohort is a good sample based on total grades and problem-solving grade. In achieving this, the EPS cohort grade average was typically, but not significantly, lower than the EPS remainder population. The EPS cohort appeared to be a fair estimate of the students who finished the year with the EPS instructor.

Conversely, the TRD cohort was an overestimate of the students in the TRD class based on total grade. In fact, there were very few C students in the TRD class who could have been eligible for the cohort. The TRD cohort also was top heavy in the tiers based on problem-solving grade. Given that the two cohorts of students started out the same, it is interesting that the TRD cohort does not look like the rest of their classmates as far as grades determine. It proved insightful to examine the predictive success of most of the cohort matching parameters to explain this puzzle.

Multiple regression for grade

The previous section showed the TRD cohort was an overestimate based on grades and the EPS cohort was a fair estimate of the students in their respective classes.

A step-wise (forward) multiple regression was executed to see which (if any) of the matching parameters could predict the total grade received. Since the total grade suggested the imbalance of the TRD cohort, the total grade will be used to ascertain if there was a difference in how the students' measured backgrounds and skills predicted their grade in each class. The students' total points were used as the dependant variable. Each of the seven matching parameters were used as independent variables in a forward stepwise multiple regression. The analysis was carried out four times: twice for the survivors in each class and twice again parsing out the cohorts from their peers. This gives a total of 6 equations.

The first analysis was done for the EPS class and included only survivor, remainder, and cohort students. There were 123 of these students who completed all the measures. For this set of students a weak, albeit statistically significant (R=0.352; F=8.465, p=0.0004) regression equation was found:

$$EPS\ Class\ Total\ Grade = 0.266 * FCI + 0.230 * MPEX + 0.573$$

This equation can only account for about 12% of the variance in the students' grades, but it does suggests that both the FCI and MPEX scores of the students had a positive effect on the grade prediction.

The next analysis was done for the EPS cohort, but this analysis split the cohort from the rest of the class. In this scenario we get two more equations. The equation for the EPS cohort is quite predictive (R=0.708; F=10.576, p=0.0007) and different from the EPS class as a whole:

$$EPS\ Cohort\ Total\ Grade = 0.582 * MPEX - 0.370 * Math + 0.513$$

In the EPS cohort the equation accounts for over half of the variance in the grades. It suggests that a student's attitude toward the class and physics in general would positively affect their grade. However, a high math preparation would hurt their grade. The FCI did not factor out into this equation.

The EPS remainders and survivors (EPS R&S) regression equation only used the students' FCI scores to produce a barely discernable regression equation (R=0.251; F=6.439, p=0.0128):

$$EPS\ R\&S\ Total\ Grade = 0.251 * FCI + 0.729$$

An interpretation from these two equations is that the EPS remainders and survivors are responsible for the FCI score in the whole class regression equation, while the EPS cohort accounts for the presence of the MPEX.

The TRD class afforded different equations. As before, this analysis includes only those students who passed all three academic terms and were enrolled in the first term class (T1) with the TRD instructor. There were 81 students who fit this description and completed all measures. The TRD class returned a moderately strong regression equation (R=0.525; F=14.870, p<0.0001):

$$TRD\ Class\ Total\ Grade = 0.485 * FCI + 0.231 * MPEX + 0.492$$

Perhaps encouragingly the equation form is very similar to the EPS class. Unlike the EPS class, the equation for the TRD class accounts for a more reasonable amount of the variance in the grades (25%).

The next two equations are produced from parsing the TRD cohort out of the TRD class. Examining the TRD cohort equation proved insightful. This equation was significant and moderately predictive (R=0.497; F=7.206, p<0.0135):

$$TRD\ Cohort\ Total\ Grade = 0.497 * Math + 0.643$$

But more important than the equation itself is its comparison with the EPS cohort. Both cohort equations used the student's math backgrounds. For the EPS cohort, the smaller of the two coefficients belonged to the student's math background - and it was negative. With the TRD cohort the only coefficient was positive for math.

For completeness, the final equation was for the TRD remainders and survivors (TRD R&S). This equation was remarkably similar to the equation found for the whole TRD class, including the correlation coefficient and probability ($R=0.555$; $F=11.998$, $p<0.0001$):

$$TRD\ R\&S\ Total\ Grade = 0.524 * FCI + 0.227 * MPEX + 0.483$$

It was interesting that the TRD matching parameters regression equations did not change between the class and the TRD remainders and survivors. In effect the removing the cohort had little impact on the analysis for the TRD class, while removing the cohort had an effect on the EPS class.

Table 4.14 provides a summary of the proceeding analysis. Of the seven matching parameters, only the FCI, MPEX and math background emerged from the analysis. Furthermore, the math background only emerged when examining the cohorts themselves. Even more provocative was the negative relationship between math background and grade in the EPS cohort, while the TRD cohort had only a positive relationship.

Table 4.14

Summary of Total Grade Multiple Regression of Matching Parameters

	EPS Course			TRD Course		
Parameter	Class	Cohort	R&S	Class	Cohort	R&S
FCI	0.266	-	0.251	0.485	-	0.524
MPEX	0.230	0.528	-	0.231	-	0.227
Math	-	-0.370	-	-	0.497	-
R	0.352	0.708	0.251	0.525	0.497	0.555

Predictiveness of SAP and LP

The last matching parameters were the problem-solving scores of Logical Progression (LP) and Specific Application of the Physics (SAP) coded from the third quiz of the first term class (T1-Q3). These were also triple weighed factors, but this high weighting proved to be inappropriate. For the EPS cohort, neither the T1-Q3-LP score ($R=0.078$, $p=0.716$) nor the T1-Q3-SAP ($R=0.129$, $p=0.548$) scores were correlated with the first term grade. This was seen in the TRD cohort as well. The T1-Q3-LP score did not correlate with the first term grades ($R=0.370$, $p=0.076$) and the T1-Q3-SAP scores did not correlate with the first term grades ($R=0.224$, $p=0.292$). From this result, the use of the Logical Progression and Specific Application of the Physics scores based on a single quiz was not a useful predictor of grade for either cohort.

Summary

From the comparisons of the matching parameters to the external referent of the first term course grades, it was evident that most of the matching parameters were not useful predictors of students success. In one case, a parameter had an inverse effect. The students' math backgrounds predicted differently between the cohorts. For the TRD cohort, math background predicted positively with grade. For the EPS cohort, math

background correlated negatively with grade. This suggests when math background was used to select EPS students based on the math backgrounds of their TRD counterparts, that the next effect was to pair-up some of the high-scoring TRD students with ultimately lower-scoring EPS students. This is evident from Figure 4.3 and 4.4. Had the math background matching parameters behaved the same between the two cohorts, the EPS cohort should have also over-represented their peers.

Implication for the TRD cohort

The purpose of this section of this chapter was to examine the third major threat to the validity of this study, namely the selection of the students to the cohorts. Recall that the purpose of the matched cohort was to create and follow two evenly matched teams of students through each of the two instructor's courses. There was an implicit assumption that the two cohorts would be representative of their fellow students in each course, or at least those students who remained with the same instructor throughout the academic year. Without meeting this assumption, issues of sampling-bias could arise. The two cohorts were compared to their classmates with their initial demographic information (the matching parameters) and overall course grade assigned to each student.

What emerged was that the two cohorts had essentially the same demographics as their classmates. However, when the grades were compared it was noticed that the TRD cohort performed better than their classmates. The EPS cohort was fair. This may have occurred due to the positive role math background played in determining the TRD student's grades, an effect observed in reverse in the EPS cohort. Whatever the cause, the imbalance remains.

If this was a statistical study, this sampling error could have nullified the results. However, this is a case-study. The non-representativeness of the TRD cohort in terms of grades essentially adds to the description of the two cohorts. Furthermore, as was discussed in the previous chapter, this study is already biased toward the TRD cohort. The selection of quiz questions and problem-solving skills were biased toward the TRD cohort. The grade mis-match of the two cohorts only further biases the study. The study can continue so long as this (and the other biases) are properly accounted for in the descriptions of the cohorts.

Usage of the Problem-solving Strategy

The final threat to the validity of the study that will be examined involved the use of the Minnesota Problem-solving Strategy by the EPS students. If the students didn't use the strategy, the interesting difference between the EPS and TRD cohorts would vanish. This section examines this threat to the study.

Usage

During the coding of the student's solutions, a subjective determination of the use of the Minnesota Problem-solving Strategy in the student's solution was made. Recall that the Minnesota Problem-solving Strategy was explicitly taught in the EPS section. The graph of usage versus time is shown in Figure 4.5. The vertical axis on this graph is the percentage of problems solved using the Minnesota Problem-Solving Strategy. The horizontal axis is the date axes. The data points are plotted consistent with the procedures adopted for development graphs.

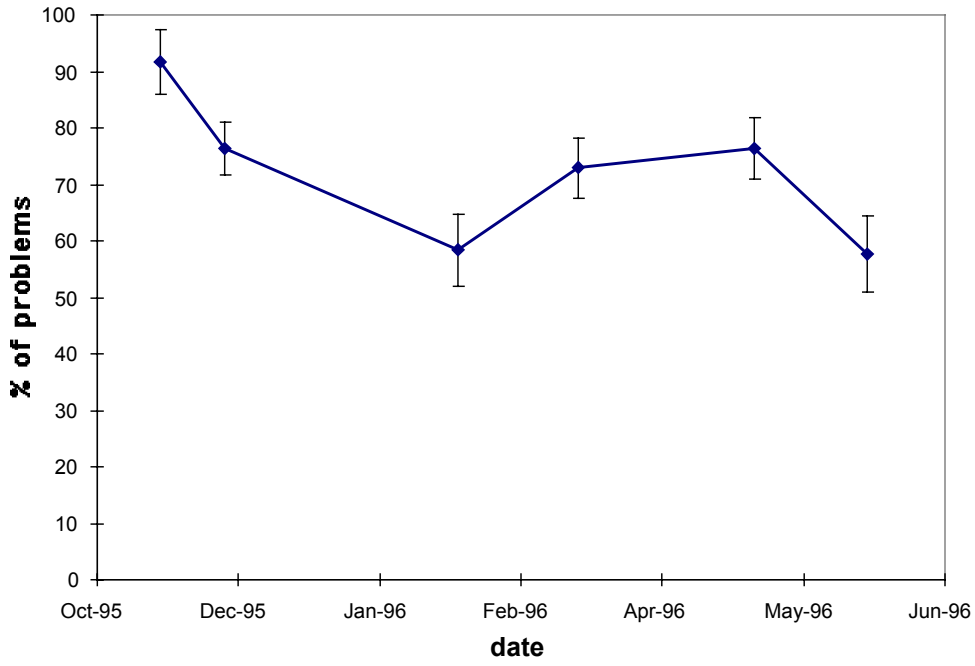

Figure 4.5
Percentage of solutions when the EPS cohort students used the Minnesota Problem-solving Strategy.

This Figure 4.5 suggested that the students in the EPS cohort generally used the explicit problem-solving strategy throughout the year for about 75% of the problems. The 60% score on the second-term quizzes is unusually low due to the first quiz of that term where only 40% of the students used the strategy to solve that problem. Perhaps the long winter break coupled with general laziness caused this one-quiz drop, which presumably was corrected by the grading. However, the drop-off seen on the third-term final was an interesting trend.

This last final exam drop-off was not caused by the problems getting too difficult. This was however a reasonable hypothesis since the difficulties of the problems do increase with time. When the usage was plotted against the difficulty rank of the problem and a correlation then computed, a significant difference was found ($\rho = -0.458, p=0.015$). However, with the confound caused by the date factored out of the correlation and the

residuals of usage and difficulty determined, no significant difference was found (ρ = -0.236, p=0.232). The interpretation of this analysis is that the problem difficulty rank and usage are not correlated. A possible explanation can be found in the MPEX data.

Maryland Physics Expectation Survey (MPEX)

The Maryland Physics Expectation Survey (MPEX) created by Redish, Saul, and Steinberg (1998) was used as a gauge of the students' general attitude of the course. The MPEX instrument can be found in Appendix E. The MPEX was administered four times during the year. The first offering was during the first week of labs. The next two offerings came at the beginning and end of the second term. The last offering was given on the last week of labs during the term. Not all the students in both cohorts completed the MPEX when it was offered the full four times. On average, 88% of the questions were answered by the EPS cohort and 85% of the questions by the TRD cohort. Even with a few missing data points, the averages for both cohorts can be computed and shown as a development graph in Figure 4.6.

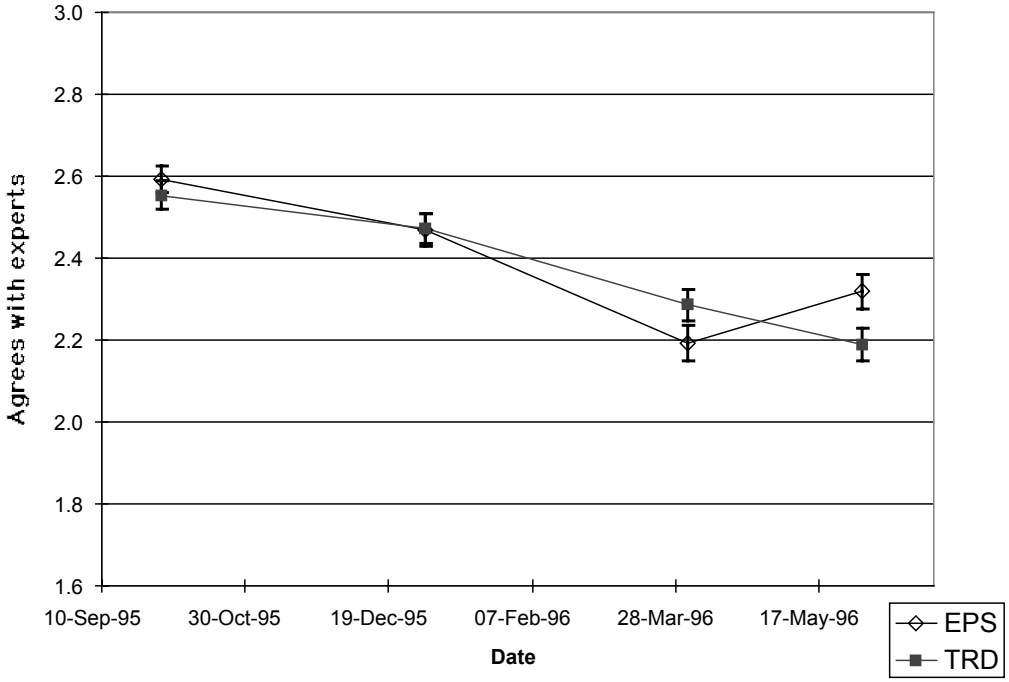

Figure 4.6

MPEX scores for both cohorts. The higher the score, the more the cohort agreed with the expert responses.

The two cohorts begin the year about the same, which they should since they were matched on the MPEX score. Both cohorts then began a gradual decline during the course of the year. The EPS cohort ends the course with a slightly less pronounced decline by the end of the year. It is this up-turn which offers a clue to the usage drop-off. Perhaps the students' increased attitude scores correspond with increased appreciation of physics; more expert-like goals. In this mind set, fewer students should need to use the strategy. While this cannot be confirmed by this study, it might be pursued in a latter study.

Conclusion

The results of this interpretative study are exposed to many validity threats. The goal of this chapter was to use data from many different sources to add validity to the results presented in the next chapter by addressing known (or predicted) threats to validity. This technique is known as triangulation and is common in most interpretive studies.

The first threat to the validity of this study involved the four problem-solving skill measurements. They were seen measure what they were designed to measure and using all four skills was important. The second threat to validity involved the use of the problem difficulty ranks to adjust the problem-solving skill development results. Again, this chapter demonstrated that the difficulty ranks are a valid and useful tool. The third major threat involved the selection of the students to the cohorts. Here it was seen that the TRD cohort got better-than-average grades and therefore over-represents their classmates. This result will need to be addressed in the next chapter. The final threat that will be examined involved the use of the Minnesota Problem-solving Strategy by the EPS students. If the students didn't use the strategy, the interesting difference between the EPS and TRD cohorts would vanish. The students generally used the strategy.

With the methods of verification chapter completed, this study can now move onto the principle results of this study: the development of student problem-solving skills.