EDUCATIONAL OUTCOMES ASSESSMENT: FRAMING& RESOLVING ILL-DEFINED PROBLEMS

Col John Russell, Vice Commander
Air Force Institute of Technology
Lt Col Rosario Nici, Department of Astronautics
Lt Col Charles Hudlin , Department of Philosophy and Fine Arts
Lt Col Donna Peterson, Department of Electric Engineering
Lt Col Steve Slate, Department of Economics and Geography
Maj Rick Abderhalden, Department of Management
Maj Mavis Compagno, Department of Computer Sciences
Dr Mary Marlino, Center for Educational Excellence
Col David Porter, Department of Behavioral Sciences and Leadership
United States Air Force Academy

This paper discusses the initial assessment of our students' ability to frame and resolve ill-defined problems. It outlines the development and administration of the assessment instrument, scoring procedures, and preliminary results. The problem used for this assessment was presented as an Air Force deployment scenario and students were given 12 minutes to develop a response. They were allowed another 12 minutes to continue and/or reflect on their thinking and problem resolution. This was followed by 15 minutes of discussion to provide feedback and closure for the students and administrators. All 436 responses were scored and demographic information, which was not available to the raters, allowed an assessment of subgroups. Preliminary investigations suggested that student performance was not affected by gender or time of day. Effects of general academic performance (i.e., GPA) or choice of major had smaller than expected correlation. Additional research to examine convergence between scores on this instrument and other more standard measures is needed.

Overview

One of the educational outcomes for graduates of the United States Air Force Academy is the ability to frame and resolve ill-defined problems (see Appendix A). Air Force Officers are confronted with problems that do not always have "approved solutions." A graduate maybe expected to address tactical employment and deployment problems, resource allocation concerns, and organizational leadership issues. Frequently, inadequate or conflicting data is all that is available. A decision which will have real consequences is required in a relatively short time. Sometimes graduates must extend simplified models and methods to the complex situations characteristic of the real world. At the same time, graduates must also recognize the limits of their simplified models, identify the significance of the

missing data and articulate their solutions with an appropriate level of confidence. Our students' ability to frame, solve and resolve problems of this type should receive increasing emphasis throughout their four years at the Academy. How well they develop this skill has always been a matter of conjecture because it is difficult to objectively and reliably evaluate. This report explains our initial attempt during AY 94-95 to assess framing and resolving ill-defined problems, describes our activities associated with this effort, and presents the results of our evaluation.

Method

Assessing cadets' ability to frame and resolve ill-defined problems is itself an ill-defined problem. The problem was framed by defining the outcome in terms of behaviors that describe different levels of performance (See Appendix B). Initially we classified performance as being excellent, satisfactory, or deficient. Next an instrument was developed and then administered to a 50°/0 sample of the Air Force Academy's Class of 1995 during the spring semester of their senior year. Student responses to the exercise were evaluated using the levels of performance.

Our task was to assess a large number of graduating seniors with an instrument that they would consider relevant. An additional goal was to provide an adequate basis upon which to draw general conclusions about the Class of 1995, as well as, specific subgroups within the class. Engineering Systems Design (Engr 410), a three-semester hour course, provided a useful mechanism for administering the assessment. All students at USAFA are required to successfully complete Engr 410. The course introduces students to the Department of Defense acquisition process through hands-on experience gained during the design, construction, testing and deployment of an engineering design project. Each section of approximately 20 students works on a different project, most of which have a public service orientation. The mix of students within each section is random, with a broad range of academic majors represented. Roughly half of the senior class takes the course in the fall semester, with the remainder completing it in the spring semester. Our assessment instrument was administered to 436 seniors enrolled in Engr 410 during the Spring '95 semester. The assessment process included a standard introduction, two-stage administration, and group debriefing for each section and was completed within single 50-minute class sessions on two consecutive days.

A crucial task was to develop the instrument itself. Although other assessment instruments have previously been developed, we felt they were not very relevant to our students, thus the students would not be motivated. Therefore, an instrument involving a relatively realistic Air Force problem was developed by members of this team. It was designed to evaluate each cadet's ability to:
1) recognize the ill-defined nature of the problem; 2) resolve the problem; and 3) articulate their solution process and level of confidence in the process outcome. The exercise was presented as a scenario that provided information pertaining to the deployment of Air Force aircraft along with historical data concerning the number of aircraft, personnel, and sortie generation of previous similar aircraft deployments (see Appendix C). The exercise asked the student to predict the success of a planned deployment, provide a level of confidence for the prediction, and provide a brief description of the process used in arriving at their "solution."

As mentioned earlier, general criteria or levels of performance had been developed to evaluate students' performance in dealing with ill-defined problems. The scenario and performance levels were tested and revised through a series of prototype administrations to individuals and groups ranging in

rank from junior officers to a retired Air Force four-star general. This process helped refine the method for administration as well as the subsequent debriefing.

The actual assessment instrument was administered to the cadets by the faculty members who had participated in its development. The lack of a script meant there were slight variations in instructions given to cadets. Students were allowed 24 minutes to complete the exercise. They were only given the first page of the assessment initially, then after approximately 12 minutes they were given the second page asking them to reflect on their own thinking and problem solving. After the cadets had completed the task, approximately 15 minutes were devoted to an in-class debriefing of the exercise and discussion of the Academy's framing and resolving educational outcome. According to the team members who administered the instrument, the vast majority of students worked diligently to accomplish the task. Students were assured that their responses would not impact their Engr 410 or any other course grade. Demographic information was obtained to analyze the influence of various factors on the Class of 1995. All individual information was removed before the evaluation process began so that raters would not know anything about subjects other than their response.

After the administration, six student responses representing all performance levels were evaluated by the Educational Outcomes Assessment Working Group using the levels of performance. This "calibration" session helped resolve level of performance interpretation issues and standardized the subsequent grading process. One of the major results of this session was to expand the number of categories from three (Excellent, Satisfactory, and Deficient), to five (Excellent, Excellent/Satisfactory, Satisfactory, Satisfactory/Deficient, and Deficient). The Working Group determined that five categories provided the evaluators with enough latitude to adequately evaluate the instrument and more readily reach consensus.

Using these five categories, we then evaluated a sufficient number of cadet responses to form an overall impression of the distribution of scores. For this evaluation 108 random responses were divided among the Working Group members. Each member evaluated a total of 18 student papers, six of which were identical for all members. The twelve additional papers were evaluated by two separate members, and any discrepancies in the evaluation of an individual response were discussed and resolved. The high level of consistency between raters validated the evaluation process and provided an initial assessment of the class of 1995. Convergence of these ratings with those of Dr. Cindy Lynch (a consultant with considerable experience and expertise in assessment with respect to the Reflective Judgment Model) further corroborated the instrument and scoring procedures.

The final step in this process was the evaluation of all 436 responses in order to assess subgroups within the class. The student responses (including those already evaluated for validation purposes) were divided among eight Working Group members. These members then solicited other faculty members within their departments to go through similar "calibration" sessions of grading a small sample of the responses. Subsequently, the results were discussed and the remainder of student responses were graded. The purpose of bringing in other faculty members was twofold: first, to broaden the number of faculty members involved in understanding and promoting the framing and resolving educational outcome; second, to distribute the workload and allow for subsequent assessment activities.

The remainder of this paper summarizes the results of this assessment and draws conclusions concerning both the Class of 1995 as a whole and subgroups identified within the class.

Results

Overall the mean score was 2.6 on a 1 to 5 scale, or approximately midway between the satisfactory (3) and satisfactory/deficient (2) categories (see Figure 1).

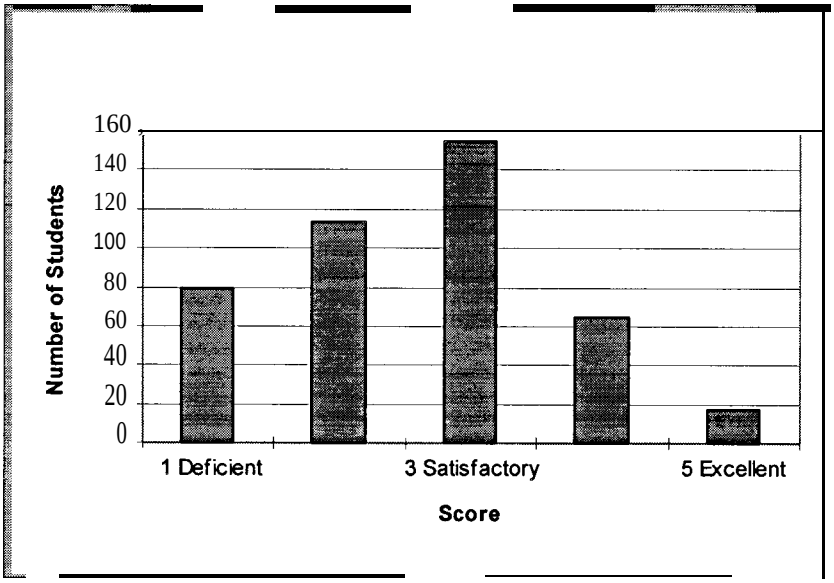

FIGURE 1
Number of Students Achieving Each Score

Even with the extensive effort at piloting, administering and scoring this instrument, it would be inappropriate to conclude that these scores necessarily reflect an objective measure of student ability to frame and resolve ill-defined problems with respect to the absolute standards listed in the levels of performance. However, 350 of 435 (80%) demonstrated some ability to frame and resolve the ill-defined problem contained in the instrument. Performance on this particular task is likely to have been affected by other factors such as effort and prior knowledge of the domain. Thus, the distribution of scores is likely to represent an underestimate of cadets' true ability (i.e., the performance expected if all of them were highly motivated and working within a knowledge domain they had studied for more than the 24 minutes allowed during this task). Nonetheless, it was not unrealistic to expect that the pattern of results across various demographic and academic subgroups might yield some insight into the educational process. From a psychometric perspective, our instrument's ability to allow raters to reliably and objectively discriminate between student responses represents an important accomplishment.

After examining the overall distribution of results, the effects of various demographic and academic factors were systematically examined. The first factor was gender. There were no reasons to expect this factor to have an effect. Next, the effects of time of day were considered. The purpose of looking at this variable was to examine whether or not there was evidence that scores changed systematically across administrators. Since students were randomly assigned to sections, no time of day effects were anticipated. The final three factors examined were expected to show some effects. The first of these was GPA. Given that most of our faculty claim to emphasize the ability to frame and resolve ill-defined problems, one might expect that academic performance (i.e., GPA) would be

positively related to performance on this task. The other two areas were academic division and academic major. One might hypothesize that students who work more with ill-defined or "fuzzy" problems would have an advantage. For example, those majoring in Social Sciences or Humanities should score higher than those majoring in the "hard" disciplines of Basic Sciences or Engineering. In fact, other studies involving Moore's Learning Environment Preference Test has shown just such a pattern of results. The final factor examined was academic major. There were 28 academic majors represented.

Effects of GENDER

The mean rating was 2.59 for males and 2.70 for females. Figure 2 illustrates that gender had little effect on how well cadets performed on this task. Actual values are given in Table 1. The slightly higher scores for females were not statistically significant ($p < .05$).

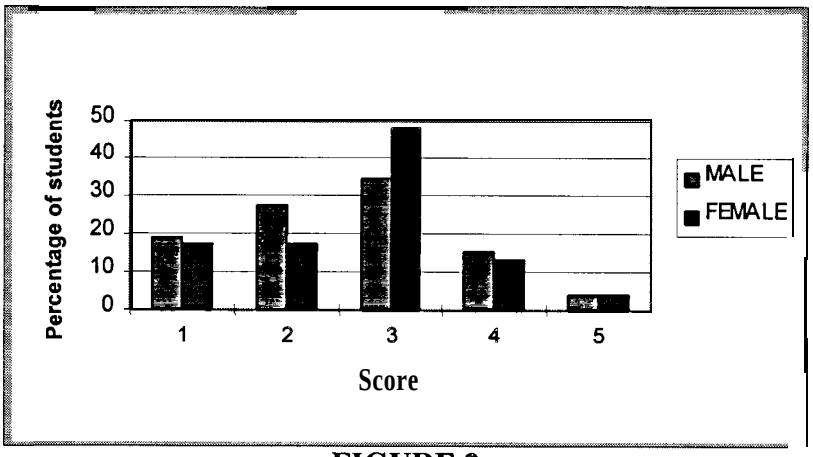

FIGURE 2
Scores by Gender

TABLE 1
Scores by Gender

	SCORE				
	1 (D)	2 (D/S)	3 (S)	4 (WE)	5 (E)
MALE	72	106	133	59	16
FEMALE	8	8	22	6	2

Effects of TIME OF DAY

Scores did not change during the course of the day. We do not know if the students told one another about the exercise. If they did, the ones with added information apparently did not ponder the problem before they got to class or if they did, it didn't change their performance. Figure 3 and Table 2 show that distributions of scores were approximately equivalent across all six meeting times. This pattern of results suggests that students' potential for handling ill-defined problems is relatively

constant throughout the academic day. There were also no significant differences observed across the two successive days in which the instruments were administered.

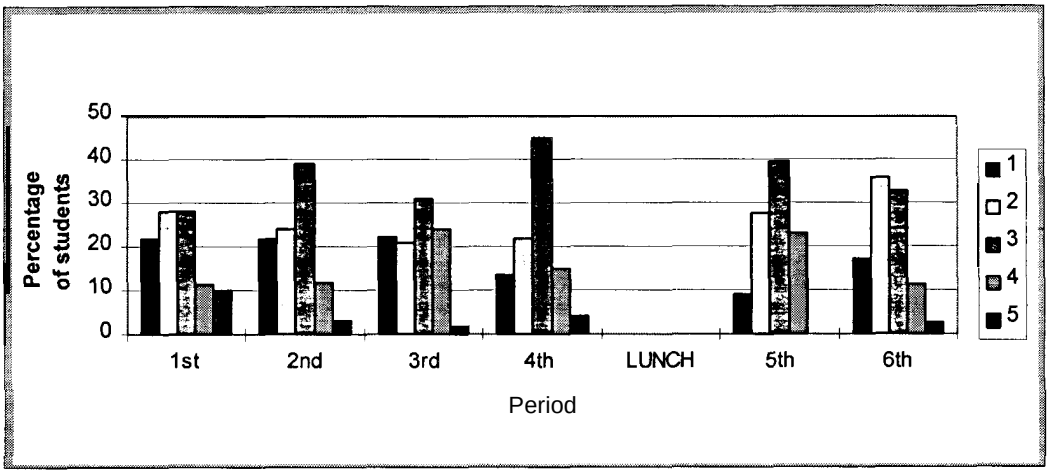

FIGURE 3
Percentage of Students in Each Period Versus Score

TABLE 2
Time of Day Mean Score and Standard Deviation

DAY ONE						
PERIOD	1	2	3	4	5	6
MEAN	2.75	2.58	3.2	2.72	N/A	2.74
SD	1.26	1.05	1.01	1.17	N/A	1.03
DAY TWO						
PERIOD	1	2	3	4	5	6
MEAN	2.44	2.38	2.42	2.76	2.77	2.3
SD	1.21	1.08	1.12	0.89	0.92	0.96

Effects of GPA

A scattergram of GPA vs score indicates that GPA is not a predictor of score achieved on our instrument, which was designed to measure critical thought process. GPA was largely unrelated to performance on this assessment instrument. Three separate measures of GPA were used: core courses GPA, overall cumulative GPA, and majors GPA. The correlation between scores and core GPA was 0.145 ($p < 0.011$) and cum CPA's correlation with performance was 0.172 ($p < 0.001$). Majors GPA yielded a correlation of 0.196 ($p < 0.001$). Although these correlation's are statistically significant, the proportion of the variance explained (viz., 2,3, and 4% respectively), is minimal. It is curious however, that majors GPA provided a much better prediction of performance than did core GPA (4% vs 2%). This previous statement spurred interest in analyzing majors and majors GPA together as an indicator of score. In general, there seem to be two alternative explanations for the lack of correlation between performance in this task and students' GPA: 1) the assessment task does not reflect students'

ability to frame and resolve ill-defined problems or 2) the assessment task is valid, but the majority of activities students are graded on generally do not involve these skills.

Effects of ACADEMIC DIVISION

Given that majors GPA was a slightly better predictor of performance than was core course GPA, one would suspect that differences in majors within academic divisions might be found. However, Figure 4 indicates there were no significant differences observed based on the academic division of cadets' selected major.

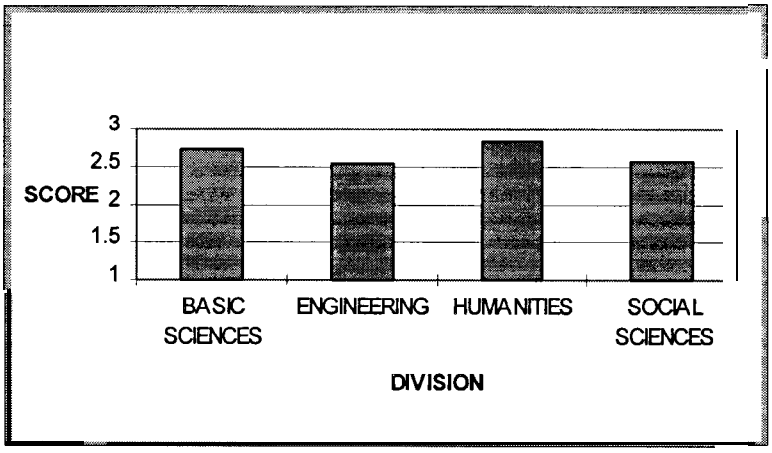

FIGURE 4
Mean Score by Academic Division

Effects of ACADEMIC MAJOR

The final analysis is an examination of the effects of individual academic majors. Although Figure 5 shows that there were some differences between majors, variations in the number of cases in each major as well as considerable variation in performance within each academic major yielded overall equivalent results. Academic major was not a predictor of performance.

Majors GPA explained only 4% of the variance for the group as a whole. Fifteen majors, (those with more than 10 respondents), were analyzed to determine if a correlation exists between majors GPA and score within that major. For example, the History majors' responses were compared with their majors' GPAs to determine if the GPA could be used as a predictor of cadet performance on this instrument. Figure 6 shows that indeed three majors had significant, positive correlation coefficients. The correlation between score and Operations Research was 0.663 ($p < 0.026$*), score and Engineering Sciences was 0.56 ($p < 0.092$), and score and General Engineering was 0.49 ($p < 0.087$'").**

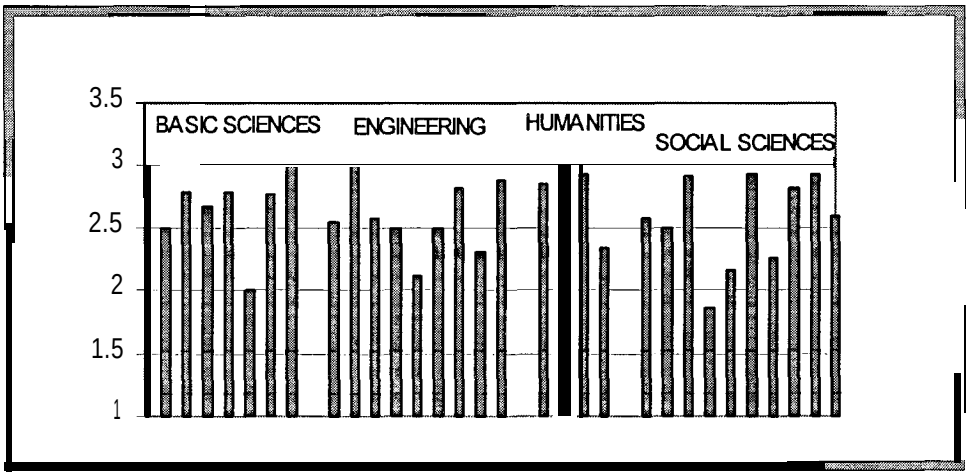

FIGURE 5
Mean Score by Major Within Academic Divisions

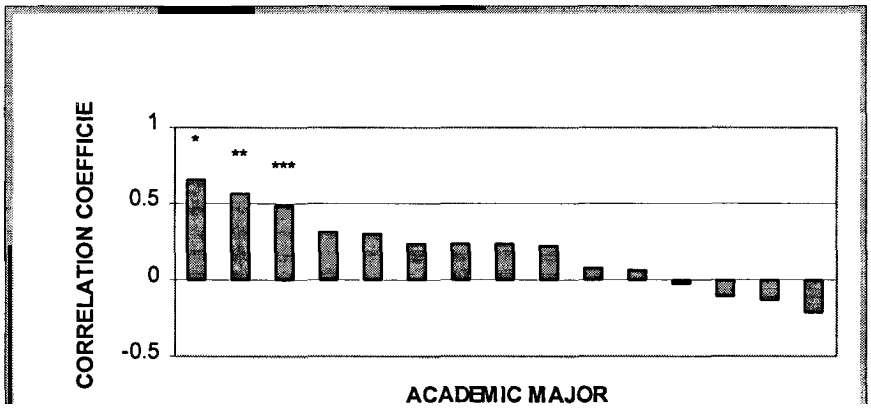

FIGURE 6
Correlation Coefficients of Majors GPA Regressed on Score

Summary

The development, administration, scoring, and analysis of responses to this instrument support several conclusions. It is possible for an interdisciplinary group of faculty to reach consensus on appropriate levels of performance for a complex skill such as framing and resolving ill-defined problems. Further, it is possible to develop an instrument, the responses to which permit reliable and consistent scoring across raters which yields a broad distribution of scores. Such a distribution suggests strong potential for the diagnostic capacity of such an instrument. Preliminary investigations suggested that students performance was not affected by gender, nor was it biased by time of day effects. However, attempts to reconcile scores on this instrument with general academic performance (i.e., GPA) or choice of academic major raise questions. Studies to examine the convergence between scores on this instrument and other more standard measures are needed.

APPENDIX A

EDUCATIONAL OUTCOME FOR FRAMING AND RESOLVING ILL-DEFINED PROBLEMS

Ill-defined problems are ambiguous, interactive and ever-changing. Framing means constructing a working model and revising it based on feedback. Resolving means that an ill-defined problem is never solved for good; rather it is solved again and again (re-solved) as the problem is framed again and again; and, each successive solution is more refined (resolution).

In assessing student skills in this area, it is important to recognize that the problem should be "ill-defined" from the student's perspective, not necessarily "ill-defined" from the perspective of experts in the field or even the faculty member evaluating performance. This suggests that different types of problems will be appropriate for assessing students general ability and their abilities within their chosen academic specialty. It is also important to point out that it is the solution process that must be assessed, not just the solution. In fact, a student who had already learned "the approved solution" from independent reading, might be less likely to demonstrate a high level of framing or resolving skills. The purpose of constructing ill-defined assessment tasks for students is to assess our graduates ability to recognize and contribute to the resolution of real world problems they are likely to face in their future careers as Air Force officers.

In addition to the definitions contained in the outcome itself, it is important to point out that ill-defined problems have no single absolute solutions. However, solutions to these problems are more than a matter of opinion or preference; viable criteria exist for evaluating solution quality. Ill-defined problems frequently contain extraneous information as well as often lack some necessary data. To provide meaningful assessment, tasks to evaluate students' skills must be carefully tailored to challenge students but not overwhelm them. Assessment of both individual and group ability to frame and resolving ill-defined problems should be undertaken across all four academic years.

APPENDIXB
LEVELS OF PERFORMANCE FOR FRAMING AND RESOLVING ILL-DEFINED PROBLEMS

EXCELLENT

- identifies most important ill-defined aspects of problem as well as general "ill-defined" problem nature
- keenly aware of personal perspective and biases and compensates effectively
- also aware of relationship between present problem and context in which it is situated
- uses goal, mission or other ultimates to structure problem space effectively
- systematically works through problem; often makes multiple passes through the problem space as conditions change in order to assesses consequences of changes or alternatives
- unsuccessful attempts regularly used to better understand problem and solution process
- generates rich variety of alternatives; tests them objectively and selects rationally
- use general principles and fundamental concepts to frame overall problem space and as solution tools; provides reasonable and substantive justification for assumptions and choices
- appropriate level of confidence and commitment to eventual solution

SATISFACTORY

- aware of general "ill-defined" nature of the problem and some of the specific problem deficiencies
- somewhat aware of personal perspective but not fully able to compensate for its effects
- evidence of awareness of problem context found throughout solution process but some important connections and implications not recognized
- may structure problem space based on superficial problem characteristics or unwarranted assumptions
- works through problem systematically but may omit necessary reconsideration of assumptions
- unsuccessful attempts recognized and abandoned
- generates multiple potential solutions but may not consider them all or use appropriate selection criteria
- tendency to use particular tools and mechanisms appropriately but may lack ability to justify the approach taken or adjust tools to fit the problem presented
- likely to lack confidence in solution; limited commitment without encouragement or support

DEFICIENT

- unaware of either general or specific characteristics that preclude routine solution procedures
- apparently unaware of personal perspectives, biases or assumptions and their effects
- apparently unaware of broader context in which problem occurs; assumes singular perspective
- unable or unwilling to structure on the problem space within parameters provided
- unsuccessful, sporadic, apparently random, attempts at problem lead to frustration and abandonment
- unsuccessful attempts based on untenable assumptions not recognized.
- fully commits to first apparent solution path and follows it through to completion without reconsideration
- random or inappropriate application of tools; may not be able to provide reasons for approach selected
- likely to display either no confidence in solution or process (may claim problem is impossible) or be inappropriately confident and overly committed to obviously ineffective solution

APPENDIXC
THE ILL-DEFINED PROBLEM

Part 1

Name _____
Section _____ **Academic Major** _____

The Commander asked me to put together a maintenance support package for our deployment of three KC-135s to a temporary operating location. When we recently deployed three aircraft, our support package had 30 maintenance personnel and we were able to fly a total of 36 sorties for the two days that we were deployed. Two years ago, we took 16 personnel and five aircraft to Eglin AFB where we flew 40 sorties in the four days we were deployed. Our sister squadron just returned from a five day trip, where they flew ten sorties in the five days they were deployed, using just five technicians and one aircraft.

The Colonel wants to fly 30 sorties in the three days we are deployed. I plan a maintenance support package that includes 20 personnel.

How effective do you think we will be during the deployment, as measured against the Commander's goal of 30 sorties? Support your position by describing how you arrived at your answer.

Part 2

Name _____

On a scale of one to ten, how do you know that your answer is correct? (absolute certainty= 10)

_____

On what do you base your level of confidence?

Does your answer depend on any particular assumptions? What are some of the most important ones?

If you could choose to have one more piece of information, what would that be?

How would that additional information change your original answer?