A Field Experimental Test of Expectancy-Valence Incentive Motivation Techniques

ROBERT D. PRITCHARD, PHILIP J. DELEO, AND CLARENCE W. VON BERGEN, JR.

Purdue University

Three types of incentive systems were developed in the context of expectancy-valence theory of work motivation. The first made small valued outcomes contingent on performance; the second attempted to make these rewards contingent on effort; the third added additional, financially-based outcomes to the reward package. These three systems were tested consecutively for two months in an Air Force technical training environment involving subjects from two training courses. The results indicated that for one outcome, the first two systems resulted in slight but meaningful increases in performance, and the third system was fairly powerful. No real performance effects were observed for the other outcome. Attitudes generally increased under the program. The results are discussed in terms of expectancy-valence theory and in terms of their practical implications. Consideration is given to those characteristics which are necessary for powerful incentive motivation programs.

The basic logic behind incentive motivation techniques is to make valued rewards contingent on performance. Such techniques have met with considerable success in a variety of contexts including training (e.g., Schmidt, Note 4), educational settings (e.g., Lipe & Jung, 1975), clinical settings (e.g., Krasner, 1971), and industry (e.g., Marbut, 1957). For more detailed reviews of this literature see Pritchard, Von Bergen, & De Leo. Note 2; Pritchard, Leonard, Von Bergen, & Kirk, Note 3.)

This research was supported by Air Force Human Resources Laboratory Contract No. AF-080-66-C-0026, Robert D. Pritchard, Principal Investigator. Requests for reprints should be sent to R. D. Pritchard, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

The research reported here is a product of the efforts of many people. My own research staff, Robert M. Mayo, Carmen M. Sandquist, Teresa Leal, Susan Ramby, Dorothy L. Hoerman, Richard Flicko, Billie Ann Bernet, and Wahl Newnam, administered numerous measures of effort. I. Thomas Birkett, Human Resources Laboratory, handled most of the efforts. My continued support of ideas and encouragement, a great deal of skill goes to the members of the Training Research Applications Branch at Chanute Air Force Base who made the system work. Capt. Joseph D. Young, Capt. Gary N. Buehner, Capt. Joseph S. Green, T. J. H. Keith, G. Wood, L. L. Swank, and Mr. George Stroud. We would also like to thank Donald Agin for his critical comments on this paper.

Philip J. De Leo is now at the Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado.

Clarence W. Von Bergen, Jr. is now at the Kellogg Company, Battle Creek, Michigan.

The theoretical side of incentive motivation techniques have been most carefully considered in the area of work motivation. A
collection of such theories have been presented under the general
expectancy-valence model (Vroom, 1964; Porter & Lawler
Biglan, 1971; Dachler & Moblie, 1973; Campbell & Priehard, 1973). The
of these theories, however, contain three general variables which do
multiply. These are valence of outcomes (rewards), performance
outcomes, instrumentality, and effort-performance expectancy.

The purpose of the present research is to utilize the expectancy-
approach to develop and evaluate incentive motivation techniques 
context of Air Force technical training. More specifically, working
the theory, three incentive systems were developed, and their effec-
tiveness and attitudinal variables assessed.

PROCEDURES AND METHODS

Overview

The research program described here consisted of two broad
components. The first phase dealt with the identification of relevant incentives
used in the program. This was accomplished primarily by content
analysis of a broad array of incentives, having been used in the past
in similar settings, and selecting for final use those deemed
attractive and feasible.

The second phase was the actual implementation of the incentive
systems in two technical training courses. Three separate incentive sys-
tems were tested. The first awarded incentives on the basis of the
performance of a student exhibited in the training course. The second
attempted to award incentives on the basis of the effort put into
the course. The third awarded incentives on the basis of effort but utilized
only the incentives used in the first two courses but also an addition of
practically-based incentives. The nature of the experimental design was such that the effects
of various incentive systems could be assessed while controlling for
other factors as curricular trends in course performance. Hawthorne effect-
variables were monitored in student ability. The effectiveness of the incentive
system was evaluated along a series of performance and attitudinal variables
as scores on exams, speed of completion of the course, frequency
course failure, frequency of remedial instruction, self and instructor
ratings of effort, student evaluations of the teaching efforts, and student satisfaction with training, and attitudes toward re-enlistment.

We shall now turn to a detailed discussion of the exact procedures and
methods used in each phase.
Incentive Identification Phase

The Preliminary incentive material

was considered absolutely essential that every attempt be made to state as completely a list of incentives as possible. The philosophy was that when material from every available source realizing that every available source would be necessary, the value of having even one additional incentive was very high.

In order to obtain this listing, various strategies were employed. For example, when the description was (Pithead, Von Boegen, & De Leo, Note 2), the main personnel were interviewed, group meetings were held with more than 80% of the instructors from the courses, a questionnaire was held at recruiting offices, and small group meetings were held with between 200 and 400 students at the base. The format of their sessions varied accordingly.

For example, one of the techniques used with students was role playing. Groups of 5-6 students were given a practice session in role playing, and then one student, the most verbally fluent group member, was chosen to play the role of a recruiting officer. A second student was chosen to play the role of a potential recruit. The recruiter's job was to provide the potential recruit with the means of joining the Air Force, and to discuss the benefits and enjoyable experiences awaiting him at the technical training base. The potential recruit was to ask questions, provide more information, etc. The remaining members of the group were to be observers and friends of the potential recruit who had gone through Air Force training.

The "recruiting" went on for approximately one-half hour. After this time, the potential recruit "returned to his friends." They were to analyze the picture, supplying information about technical training, etc. The student failed to mention. The rationale for this procedure was that a recruiter should receive positive things that could potentially be used as incentives while the main's friends should supply negative things.

Division of Incentives for Further Study

At this point, a large amount of raw material had been generated. It was then filtered from complaints or those with little utility. Incidents were written and reviewed, duplicated, dropped, items unsuitable, etc. This procedure resulted in a list of approximately 50 items, potentially usable incentives.

Evaluating the Attractiveness of the Incentives

A critical aspect of any incentive motivation program is utilizing incentives which have high value or utility to the people in the system. This, it was
important to evaluate the attractiveness of the incentives to be used. To measure attractiveness of the incentives, a questionnaire was used. This questionnaire contained attractiveness two ways. One method was employed at the 17-point Likert scale ranging from “Extremely attractive” is one of the very best things I can imagine happening to me,” to “Neither attractive nor unattractive, I would not care whether happened or not,” to “Extremely unattractive.” It is one of the very best things I can imagine happening to me.” Additional verbal scale occurred in every other point in the 17-point scale.

The second method of measurement consisted of having the air provide a dollar value or each of the incentives. They were instructed in the Air Force had given them an extra $100 which could be spent to purchase the incentives. It could not be saved. Any or if the $100 could be spent on each incentive. Moreover, the $100 could be used over and over again until they did not want to be split up over the extra desirable incentives.

Order of presentation of the two methods of measurement was varied between so that half the respondents received the Likert scale followed by the money scale, while the other half received the reverse order of presentation. All in all, questionnaires were received from 264 subjects in either courses, 2 students, three, students in other courses.

In order to select a manageable number of maximally attractive incentives, decision rules were generated for picking items with positive attractiveness or high negative attractiveness. It was decided to group those items which had a mean Likert value of +4 or greater, or —2 or less, and a mean Money value of $30.00 or more, or less than 30.00. It was planned to ultimately devise a system whereby writers would be able to choose incentives from it. For example, anatomic’s superior performance could be chosen to be excluded from training that got a 3-day pass. Thus, the mean attractiveness score was the criterion for selection. It was felt that even if an incentive was a moderately high in mean attractiveness, but had a great deal of variance, it should be selected. That is, although many students saw it as an attractive incentive, many others saw it as high. Then, for the better good would serve as a powerful incentive. Consequently, a third criterion was adopted. If an incentive had a mean Likert value greater than +3 less than —3, and had a standard deviation greater than 5, it was selected.

This procedure resulted in the selection of approximately 20 incentives.
**Incentive Motivation Techniques**

...only very superficial attempts had been made at this point to remove mental incentives on grounds of low feasibility. The next step, then, was carefully to explore the feasibility of the remaining incentives. To accomplish this, a series of meetings was arranged with those people who might be directly or indirectly affected by the system.

Based on the comments and reactions from these meetings, many of the incentives had to be dropped. This elimination occurred for a variety of reasons. For example, as we had expected, use of certain incentives could not be approved at Base level, but required higher approval. Such approval was sought, but in most cases was not forthcoming for various administrative and policy reasons. Other reasons for eliminations included lack of needed funds, difficulty of administration, or conflict with Air Force regulations.

**Table 1**

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Attraction of Incentives (208)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short scale</td>
</tr>
<tr>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1. Special award for outstanding performance</td>
<td>5.76</td>
</tr>
<tr>
<td>2. Special award for outstanding performance</td>
<td>5.87</td>
</tr>
<tr>
<td>3. Being excused from squadron details for a week</td>
<td>4.31</td>
</tr>
<tr>
<td>4. Getting a free pass over the weekend</td>
<td>5.64</td>
</tr>
<tr>
<td>5. Getting a Walker's Pass for a week</td>
<td>4.28</td>
</tr>
<tr>
<td>6. Being able to wear coveralls in the Order</td>
<td>5.33</td>
</tr>
<tr>
<td>7. Not having to go to classes for 1 day</td>
<td>3.26</td>
</tr>
<tr>
<td>8. Wearing any uniform desired for a week</td>
<td>3.60</td>
</tr>
</tbody>
</table>

---

*Note: Scales are on a 0-9 scale. Positive scale is incentive; negative scale is detractor.*
The Experiment

The second phase of research consisted of taking those students identified as both powerful and flexible and making them contort, in various ways, on performance in the two target courses.

The Setting

The experiment took place at Chicago Air Force Base, a training base located in Ramrod, Bermuda. The bulk of these students were instructor men who had come directly from basic training at Lackland Air Force Base. Two courses at the base were selected for this research, the Aircraft Electrical Repairman course and the Weather Observer course.

The Aircraft Electrical Repairman course (AER) was a self-paced course utilizing programmed texts and was originally planned to be six weeks in length. Since it was a self-paced course, however, the time to complete it was variable. The purpose of the course was to train students to perform the duties of the Aircraft Electrical Repairman. These duties include inspection, maintenance, and minor repair of electrical systems and components in aircraft and associated equipment.

The course was organized into six parts or blocks. The student took each block sequentially. At the end of each block, a student had to take and pass a block exam. This exam was taken when the student had completed all the programmed texts and other materials in that section of the course, and felt that he was prepared to pass the test. Each block exam consisted of a written, multiple-choice test on which the student had to complete a score of 80 to 85 (depending on the block) to pass. In some of the blocks a student also had to pass a performance test which required him to actually perform some operation on a piece of equipment. He was scored either satisfactory or unsatisfactory on his performance test by the instructor. Students who failed to pass any part of the block exam were required to retake it. After failure of a block exam, a student could be required to attend remedial instruction sessions and receive counseling by his technical instructor. A student was required to attend remedial instruction session whenever, in the judgment of his instructor, he needed such instruction. These sessions were in addition to the students' normal class day. A student's final grade in the course was the numerical average of all his written block examination scores.

The number of students in the AER course at any given time varied from about 200 to 400. The students were 20 males, and most were between 17-20 years old. Most had finished high school, and a small number had had some college. The entrance requirements for the course were relatively low: 40th percentile or better on either the Electrical or Mechanical subtests of the Armed Qualifying Examination (AQE). Thus, the students in the AER course tended to be of lower ability than
In many of the other courses at Chanute, and quite a bit lower in rigidity than those in the other target course, the Weather Observer (WX) course.

The other course that was selected for study was the Weather Observer course. This course was a lock-step course 16 weeks long. In contrast to the AER course, the lock-step nature of the course with its corresponding calendar instruction did not allow for variations in the speed of completing the course.

The purpose of the WX course is to train students in the areas of weather observation, plotting weather maps and charts, meteorology, and weather equipment operations.

The course was organized into four blocks. The evaluation system was somewhat different than that in the AER course. Students took a series of multiple-choice tests throughout each block. The average of those tests, known as "measurements," constituted the block score, and the average of the block scores constituted the final course grade. Students took 20 measurements during the 16-week course. A student had to pass each measurement. In cases of repeated failures of a given measurement or failure of several different measurements in a short period of time a student could be "washed back." This meant that the student had to repeat the previous 2 weeks of the course. As with the AER course, a student whose performance was poor could be required to take sessions of remedial instruction and receive counseling.

The number of students in the WX course varied from about 150 to 18. The students were predominantly male, but there were some female students. Most of the students were between 17-20 years old, almost all of whom had finished high school, and a good proportion had had some college. The enrollment in the course required a minimum score of the 80th percentile on the General Subtest of the Airman Qualifying Examination (AQE) the students in the WX course tended to be of fairly high ability, substantially higher that the students in the AER course.

A new class entered each of the two target courses every 2 weeks. At any given time students were at all different points in the course. Students in any two target courses were in the same administrative unit (squadron), and no other courses were represented in that squadron. Students spent 6 1/2 days, 5 days a week in class. In the morning shift ("A Shift") classes went from 6:00 am to noon; the afternoon shift ("B Shift") from noon to 6:00 pm.

Students on "A" shift would get up at approximately 5:00 am, march breakfast, eat, and march to class by 5:50 am. They would attend classes for 6 hr, with breaks every hour, and march to lunch. After lunch they were free until early afternoon at which time they were required to attend "Tl Notes." The purpose of the "Tl Notes" formation was to make...
Personnel, assign work details when applicable, etc. Students did not have work details, medical appointments, drivers training classes, remedial instruction, etc., were told not to do so as they wished, including leave the base. On any given day most of the students did have free time during this period. Students were also free to make the base on weekdays except when they were assigned "duty" over the weekend. However, any given student was not infrequent.

Students on "D" shift had a similar schedule, except for "D" Notes, may round 0300 rm and they used to close from noon to 0600 rm.

Dependent Measures

The dependent measures for the research program consisted primarily of performance-based measures and attendance measures. These will be discussed in detail below.

Most sources of performance data came from informants conscripted on a sheet that was kept by the technical instructor for each student. This information included block exam scores, final course grade, length of time to complete the course, exam failure, remedial instruction and examining sessions.

The other source of performance data, comes from instrument developed especially for this project. It was felt that in addition to usual performance, measure of effort would be valuable sources of additional data. Consequently, a measure of effort was devised. Based on interviews with technical instructors and intuitive analysis, a set of tests was generated which tapped various aspects of effort in technical training. These items dealt with such things as the level of energy expended by the student, frequency of asking questions in class, level of effort in class, frequency of staying outside class time one of class hours, etc. Five items were used plus one item measuring overall effort. Each item was expressed in a 0–10 point Likert format.

Two types of the measure were used. One form was for students at the completion of each block of the course. This seemed reasonable since each instructor typically had to more than 20 students in a given class. One point that must be considered is that the instrument did not assess the actual performance of the students they were rating. It is quite likely that this information had some contributing effect on their ratings of effort. However, the student's knowledge of his own performance could have contributed his self-ratings of effort.

The second major class of dependent measures dealt with measures of
...attitudes. During the last week of the course each student was a short (half-hour) battery of attitude questionnaires, it had been ideal for some time to use one questionnaire which was a student critique of the course at the end of the course. The answers discussed below were merely included when this questionnary was administered... first questionnaire was titled the Student Opinion Questionnaire. It by consisted of three separate measures. However, the items for the measures were mixed throughout the entire questionnaire. The first one was adapted from an experimental version (Hedeen, Note 1) of a violent crime already in use. It included a total of 31 five-point Likert items dealing with the quality and adequacy of interaction, don't help, training methods, training location, visual aids, training content, time, and physical classroom conditions, etc. A second measure embedded in this questionnaire dealt with overall satisfaction with the course. It consisted of seven items in a five-point Likert format. The items dealt with such things as the efficiency of the Force, the Air Force's concern over the individual, living conditions in the Air Force, the importance of the Air Force's mission, and military attitudes related to the Vietnam war. The third measure in the Student Opinion Questionnaire was a measure of desirability adapted from the Crowne-Marlow Social Desirability Scale (Crowne & Marlow, 1960). Ten items in a five-point Likert format were used, all of which were reversed to fit the theoretical training... The next questionnaire was designed to measure the job satisfaction the students felt in their position in the Air Force training. It was adapted from Minnesota Satisfaction Questionnaire (Witta, Davis, England, & Pata, 1967). The measure contained 22 items in a five-point Likert format. The items dealt with satisfaction towards such aspects of the training task as independence, variety, supervision, security, responsibilities, fairness of Air Force policies, pay, peers, and feeling of ... The final questionnaire in the battery contained five items in a nine-point Likert format dealing with attitudes toward volunteering. It included a short volunteering for the Air Force, leading for a second term, attitudes toward a career in the Air Force... Two sets of responses were collected from all of the instructors...
project. One was a check on the experimenter's manipulations and the other
with students' and instructors' attitudes toward each of the four
incentive systems. These will be discussed in a later section.

Major Experimental Conditions

Three types of incentive systems were tested in this program. They
were run sequentially over a period of approximately 8 months. As a "bag
overview, the first system gave incentives which could be earned only
by those facilities and resources available at the base itself. These
incentives were awarded to the students on the basis of their normal
performance in the target courses. The second system utilized the same
incentives as the first system, but awarded them differently. A novelty
was devised which took into account a student's level of ability and, in
existence, gave students handbooks on the basis of ability. In theory, the
rewards in amount of incentives on the basis of effort. The third incentive
system utilized the textbook system, and the incentives used in the first
systems, but added additional incentives in the form of financial
based incentives.

Incentive system No. 1. The first incentive system was a "classical"
system in that valued rewards were made contingent on performance; sol-
the higher the performance, the more the rewards. Two other "classical
systems were considered, but rejected. First, a system whereby a panel
gets some reward if its performance surpasses some "cutoff" point from
systems. This has the disadvantage that there is no incentive for performance
beyond the cutoff point. A second type of system's in where the top
performing group (e.g., top 25%) gets rewards. Such a system could encourage students from helping each other since it's a
student helps another student, it benefits two students in by being in the top group.
From an expectations-value point of view, both of these two systems increase performance-reward instrumentality, but they do not maximize it.

The system actually used does maximize it.

The second system used was designed so that students received pay
for their technical performance. The higher the performance the greater
the number of points they would earn. After earning points, students
could then select the incentive they wished to have by "buying" it with
those points. The incentives varied in cost as a function of their value to
the students and their feasibility of administration. The idea of this first incentive system can best be presented in a
series of issues that had to be considered in designing the system. These
issues were: What incentives would be used? How much should they cost?
What behaviors should be rewarded and how much reward should be
given? How should the mechanics be handled?

The first issue, the incentives to be used, was based on the first phase
The research. After the process of identifying valuable and feasible incentives, a list of eight incentives remained. These incentives were refined and endorsed below.

**Commandant certifies** that the 8th Instructor at a new base is completed. Since the new base of assignment for this student was one before the student visited, the new residence was possible. The student decided to have the residence fees at the new base.

**Commandant certifies** that the new resident at the new base is certified. The student certified the student to the new base as the student actually did. The student was formally typed and signed the note (Chief in Full Color).

**Commandant certifies** that the student is certified. This certification was similar to the student at the new Commandant Officer. Both of these certifications are sent only if the student indicated that he wished to be.

**Walker's Pass for one week.** Under normal circumstances, students in a medical establishment, marched to breakfast and marched to lunch, marched back from lunch to the medical area. This incentive enabled the student to walk to these areas without any special permission.

**Being able to leave class 10 minutes.** This incentive could be purchased through a fee of $1, $2, $3, $4, etc. This incentive, as well as the others, involves time off from class, presented as a payment to the instructor for the additional time off.

**First year students** had an approved any time off from the course. If a student requested time off during a part of the course he considered final, he could require the student to excuse another time to take off. However, this restriction was rarely used.

**Getting a day off during the week.** This incentive allowed the student to get off class on during the week, Wednesday, Thursday, or Thursday. A student could get a Monday or Friday off with no incentive since that would entail a 2-day pass, a more expensive incentive. Also, he could not combine a day off with a 2-day pass and get, e.g., Thursday through Sunday off. This would be against the Air Force regulations since it would result in more than 72 consecutive days off.

**Getting a 7 day pass.** This incentive was essentially getting a Monday or Friday off with no responsibilities over the weekend. Thus, a student could be off the base for 72 hours.

**Wearing any uniform for class for 1 week.** Under normal circumstances, uniforms were mandatory for class. These are usual, tight-fitting uniforms. Several other uniforms are available, all of which are generally considered more attractive than the uniforms. Tints and this incentive allowed students to wear any Air Force approved uniform to class, if they wished. The only concession to this was when the weather was very cold in which case the heavier uniform jackets had to be worn.
Being carried from identical details for 1 week. This incentive enables
the student to be excused from such details as cutting grass, mowing
lawns, picking up debris, etc. Unfortunately, all squadron details could
be included because the squadron supervisor's personal felt there must
not be enough students to do certain details if they had been taking
these details included. General, weakened barracks guards, and prepare
for major inspections. However, for any individual details these details
were very infrequent.

Many of the instructors, both military and technical, felt that the
incentive system as it was originally proposed took away much of the
power to attain discipline, discipline, discipline for performance, etc. A frequent
cried example was the student who was doing well on his tests but in
attacking class, was a discipline problem in the squadron area, etc. The
base personnel felt that this type of person should not receive such
privileges.

To satisfy their concern, an element was built into the incentive system
that allowed an instructor to deauthorize a student from taking any privilege
from a specified period of time, usually 1 week. The student would lose points he had earned previously and could earn more points during
this "turn-off" period, but he could not use any of the actual incentives.

The second major issue in the design of this first incentive system was
how much such incentive should cost. Recall that it was decided to use
an incentive whereby the students could choose their own incentives
from a list of incentives. This technique has the advantage of maximizing
the value of high performance for each student, since the students' interests are
always to a certain extent of their own choosing rather than imposed by
the system. Another advantage is that in such a system there is a considerable variability in the cost of each incentive. The student then has
the option of potentially purchasing an inexpensive incentive or saving
his points for one more expensive.

The actual setting of the cost of each incentive was based on the
criteria read attractiveness by the students and feasibility of administra-
tion. The first consideration was that the incentives which were very attractive administratively (i.e., price was raised somewhat, and for those incentives which were very cost administratively (i.e., Walker's Pack and choice of uniform) the price was lowered. These considerations led to assignment of the following costs to
each incentive:

<table>
<thead>
<tr>
<th>Incentive</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-Day pass</td>
<td>30 points</td>
</tr>
<tr>
<td>Day off class</td>
<td>25 points</td>
</tr>
<tr>
<td>One hour off</td>
<td>3 points</td>
</tr>
</tbody>
</table>
(2) The higher the performance, the higher should be the level of rewards.

(3) Maximum possible performance (i.e., 99th percentile) should earn the incentives in the system.

With these decision rules in mind, it was decided that students would start earning points (to be used to buy incentives) if their performance was at the median of students' performance throughout the present year. Thus, 80% of students in the experimental system failed to increase performance over the level of the past year, 50% of the students would acquire at least some points. However, performance at this 50th percentile level would result in very few points. Furthermore, students whose performance was near the maximum possible (i.e., 99th percentile) earned enough points to be able to purchase all the available incentives on a regular basis.

The final set of issues in designing the first incentive system dealt with how the mechanics of the system would be handled. Basically the system required that performance be translated into incentives. To do this, it was necessary that information on performance be obtained, this information be translated into the number of points earned, these points be given to the students, the students selected the incentives they desired from the list available, and that the students actually "consume" the incentives.

Once a student had taken an examination his score was reported and the points he had earned were calculated and added to his "account." Once he had accumulated enough points he bought incentives by completing a form indicating the incentives he wished. He was then given an authorization slip for the incentive.

These mechanics of the system may seem quite complex, but in fact understood, they functioned quite smoothly and rapidly. For example, a student could typically be using an incentive from 4th to 6th after he had taken an exam.

All of this information about the mechanics of the system: incentives, cost-performance-point contingencies: suspension of privileges: logistics of getting incentives, was explained to the students and instructors in a series of briefings. In addition, all students and instructors were supplied with a manual describing the system in detail.

Incentive system No. 2. The second of the three incentive systems was similar to the first in that the same incentives were used, and the same mechanics employed. The difference was the manner in which incentives were given. In the first system incentives (actually points) were given at the basis of raw, observable performance. In the second system a type of handicap system was used which considered the ability of the students. Theoretically, this system gave rewards on the basis of effort.

Such a system is desirable on both theoretical and practical grounds.
In the theoretical side, one could argue that test performance is primarily a function of two classes of variables: ability variables and motivational variables. Furthermore, in a setting such as Air Force technical training as well as many others, it is impractical to consider raising levels of all abilities: intelligence, motor control, clinical ability, etc. Thus, to increase performance, one must increase motivation. In fact, this was exactly the purpose of this whole experimental incentive program.

Theoretical models which deal with task motivation such as those proposed by Vroom (1964), Porter and Lawler (1968), Campbell, Hackett, Lawver and Weiss (1970), and Lawler (1971) imply that to increase motivation (i.e., increase effort) one must (a) make valued outcomes contingent on high performance, and (b) make high performance contingent on high effort. A classical incentive system such as that employed in our first incentive system satisfies the first condition: high grades for high performance. However, individuals in such a system may not perceive that their level of effort is related to their level of performance.

This could occur in several ways. Individuals may not receive clear enough feedback about their level of performance, or there may be additional constraints which limit the relationship between effort and performance (e.g., an assembly line) or effort may be expanded on tasks that the organization does not consider important, or some individuals of low ability may feel that even their hardest effort will not result in high performance. It is this last situation which is probably most relevant in Air Force technical training. Specifically, low ability students may see the task as difficult and not impossible to be a high performer, and thus get the rewards of the system. Thus, the low ability student would not be influenced by a classical incentive system.

The implication of this line of reasoning is that incentives should be given on the basis of effort rather than on the basis of performance. Such a system would maximize the relationship between effort and rewards and presumably, increase effort.

To accomplish this, however, it is necessary to have some measure of effort. This is indeed a difficult situation. However, a technique was devised for this project which attempted to do this. It started with the assumption that performance is largely a function of ability and motivation. With this assumption, if one has measures of both ability and performance, one can devise a measure of motivation (i.e., effort). Specifically, if one collects a sample of correct performance data from students who have completed the course as well as the ability test data available for these students, one can generate equations through multiple regression to predict course performance from knowledge of ability data. With such an equation it is possible to examine performance as a function
of ability and motivation to make statements about motivation, I now merely subtracts predicted performance from actual performance to get a measure which should be related to effort.

Thus, if one were to predict for each student his level of expected performance from the ability-based regression equation and give incentives as his performance went above his predicted level one would be giving incentives on the basis of effort. The high ability student would have a high, "target score" and the low ability student would have a low, "target score" but it would be equally easy for both to admit or surpass the respective target scores.

Using this theoretical line of reasoning, adapted primarily for several assumptions. For example, it is a certainty that paralleling our ability for performance leaves variance over and above that strictly to motives. At the very least, however, such a technique provides for equal, people, on the basis of ability, simply giving incentives on the basis of increases in performance over the level of predicted solely in the basis of ability. The major advantage, of course, of such a system, is that it enables the lower ability student to have an equal chance of earning incentives since he can earn incentives by high effort, even though he may not be a high performer.

Such a system was the basis for the second incentive system. Ability and performance data were collected for students who had completed the courses during the 1 year preceding the start of the incentive program. Ability data consisted of the Armed Forces Qualifying Test (AFQT), general aptitude test, and the ARMN Qualifying Examination (AQE) made up of four scales: general aptitude, administrative aptitude, mechanical aptitude, and electrical aptitude.

These ability data were used as predictors in a stepwise multiple regression to predict final course grade in the Weather course and final course grade and total time in course for the Electrical course. The development and cross-validated multiple $R^2$ are listed in Table 2.

These equations were then used to assign "target" black exam scores for the WX students and "target" black exam scores and completion speeds for the AER students. The predicted scores, or "target" score represented the point at which the student started doing points. The score he went above the target score, the more points he earned.

The transition from the first to the second incentive system was handled by giving the students new manuals and by briefings. Also, each student received his own personal target score. They were told that the new system would go into effect 5 days after the briefings and that points that had already accumulated would not be taken away, they could spend them as before. The actual incentives were unchanged as were the "prices."
Incentive Motivation Techniques

TABLE 3

<table>
<thead>
<tr>
<th>Grade</th>
<th>A1 &amp; A2</th>
<th>R</th>
<th>Cross-validated R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senior</td>
<td>ARIE, Administrative</td>
<td>.53 (N=62)</td>
<td>.52 (N=42)</td>
</tr>
<tr>
<td>Junior</td>
<td>ARIE, General, ARIE, Medical</td>
<td>.59 (N=560)</td>
<td>.58 (N=305)</td>
</tr>
<tr>
<td>Freshman</td>
<td>ARIE, Mechanical, ARIE, General, Administrative, ARIE</td>
<td>.52 (N=580) (insufficient data for cross-validation)</td>
<td></td>
</tr>
</tbody>
</table>
student at a given performance level in the third incentive system receive all the incentives (points) he would have received under the second system plus additional points for the financially-based incentives. The system was designed so that the financially-based points were worth the equivalent of 75 cents. At this level it was quite possible for students to earn in extra $45 worth of incentives each week, which some students did.

The actual financial incentives and their cost are presented below.

Class for IX. Facilitators. These facilities included merchandise and services such as clothes, uniforms, personal hygiene items, current sports equipment, beer, liquor, beauty products, and other non-food items, pizza, etc. The students were issued printed certificates redeemable for any of these items. Each certificate was worth $1.00 and they had to be "purchased" in units of $300. Thus, $3.00 worth of certificates cost the student four Financial Privilege Points.

United States Savings Bonds. A $25.00 savings bond could be purchased for 25 points. It could be redeemed for each (1875) after 60 days or could be held to the student's interest.

Gift certificates for Sears. Gift certificates for Sears were issued in multiples of $6.00 (8 points each). A Sears catalogue store was located approximately 1/2 mile from the base, and a regular store was available about 12 miles from the base.

Class at the Airman's Club. These were purchased in multiples of $3.00 (4 points each) and could be used for beer, mixed drinks, etc., at the Airman's Club located on the base.

Round trip bus transportation. Students could purchase round trip bus tickets to anywhere they wished. The cost in points was determined by the cost of the ticket (1 point = 50c).

In addition, the other incentives, each of these, could be purchased at any time as the student wished.

The transition from the second incentive system to the third system was handled in a similar fashion to the previous transition. Students were informed of the new system and given new manuals. Points earned during system 2 were still negotiable for base privileges in system 3.

As the end of the third incentive system the students were informed that the entire incentive system was about to end (in 5 days) and that most of their accumulated points would be lost at that time.

In order to assess the effectiveness of the incentive system it was necessary to establish a baseline which was not exposed to the incentive system but which was relevant for comparison. Actually, data were collected for several types of baselines. The first type of control data consisted of measures of course performance in the two target courses for the year immediately preceding the start of the program. These data were...
In the present study, performance data were collected in order to assess any possible trends over time. Specifically, it was possible that performance in the courses could be higher or lower at different times of the year.

However, the data available on this cyclical control group consisted of performance data. Most of the attitude data were derived from questionnaires designed specifically for this project. Thus, a baseline for new attitude measures was also necessary. To generate such a baseline, attitude questionnaires were administered to the students in the target course throughout the period when the information on potential incentives was being collected. (This information on potential incentives was obtained at the time from students in courses other than the target course.) These attitude questionnaires were administered when the student graduated along with questionnaires that were normally given at the time by the personnel. As such, they consisted of no real change in wording for the students.

Thus, both performance data and attitude data were collected, forming a standard control group where no intervention by the research staff occurred.

In addition to this standard control group it was necessary to obtain data on a possible Hawthorne effect. A Hawthorne effect could occur in this study since the incentive systems definitely constituted special treatment. If such an effect were to occur when the incentive conditions were started, it would artificially inflate any positive effects on performance and attitudes due to the incentive system itself.

To assess the presence of such a Hawthorne effect another control condition was generated. Since the basis of the Hawthorne effect is special treatment and attention, giving the students special treatment and attention in the absence of any incentive system should produce any Hawthorne effect that was going to materialize. Furthermore, any changes in performance and attitudes due to the special treatment could be detected and thus partitioned out of the incentive effects.

To accomplish this, several types of special attention were generated. First, the students in the two courses were moved together into the same location and worked as a class. They had been together at one time, but were separated for administrative reasons. At the start of this Hawthorne condition they were moved back together. Second, they were told that they were to participate in a special research study being conducted by the Air Force and Purdue University. Finally, they were given many questionnaires, participated in group interviews, and were generally led to believe that they were involved in a special project.

Pleasing to the start of this Hawthorne condition they had had no real contact with the research team or research effort. As stated before, they were given our special questionnaires as they graduated from the course.
Walker's first \[ \text{3 points} \] week
Out of details for the first \[ \text{3 points} \] week
Choice of unifiers for first \[ \text{2 points} \] week

The other incentives, certificates, and points were handled differently. It was felt that it would not be appropriate for the students to lose these incentives since the obtaining of such a certificate would then be at least partially a function of whether a given choice to purchase it rather than select a point. Thus, if two high-performing students were assigned to the same place, and one of the two chose to "buy" the letter, the new Commanding Officer might get the impression that the one without the letter was a poorer performer. Consequently, it was decided to award these certificates if a student's performance was at or above the 50th percentile on each test made up from the performance data for the past year.

The third stage was in designing the first incentive system, what behaviors to reward and how much reward to give. It was obvious that the major type of behavior to be measured by the system was technical school performance. In the Weather course, this was fairly straightforward since the major performance variable consisted of scores on the examinations. Other performance measures such as probation and general inspection were also available, but since these were all indicators of poor performance, they were not an appropriate basis for giving positive incentives. Thus, for the Weather course, positive incentives were made contingent on scores the students received on the examinations they took throughout the course.

The situation for the Electrical course was a bit more involved since both scores on exams and speed of learning this self-paced course were relevant variables for positive incentives. In discussions with the course supervisor, it became clear that while both were important, they felt that the speed of completion criterion was more important than the scores on exams criterion. Consequently, both were used in the system with emphasis on the speed of completion criterion.

An important decision in the design of the system was how much reward should be given for a specified level of performance. This is an obvious issue since giving too much reward (too many incentives) for a given level of performance would encourage higher performance; and giving too few rewards would not make them worth working for. To arrive at the actual performance-reward contingencies, the following decision rules were used:

1. It should be possible for a large proportion of the students to get at least some incentives.
but they had no reason to think that these questionnaires were any different from the standard questionnaires that had been mailed to administrators to other graduates long before they themselves ever entered at the basic. There had been extensive interviewing and questionnaires administration by the research staff previous to this Hawthorne condition, but at least it had intentionally been done with students in other courses where place of installation and location of barracks was physically near from the students in the target courses.

After a sufficient body of data had been collected for the Hawthorne control condition, a period of eleven to six months elapsed while the mechanisms of the first incentive system were being worked out. It was felt feasible to continue the Hawthorne condition during this time due to the manpower required to give the special attention.

At the end of March, 1972, the first incentive system was started. A described previously, the three incentive systems were run simultaneously.

A body of performance data was also collected after the last incentive system to assess the after-effects of the incentive program.

Finally, a set of performance data was collected from a sample course not in the program. These data were collected during the first periods of the various experimental conditions to detect any residual changes in performance.

The various aspects of the experimental design are presented schematically, along with their scheduling in Fig. 1.

RESULTS

Overview

We now turn to the analysis of the data collected in the program. A great deal of data will be presented, an overview of the organization of this section is appropriate. We shall first consider some preliminary points which will aid in the interpretation of the results. We shall then turn to the effects of the program on performance, discussing the selection of performance measures.

![Diagram of experimental design](image-url)
Determinations of Appropriate Baselines

It is essential in a field experiment such as this to select an appropriate baseline with which to compare the effects of the experimental program. In order to do this, several sets of performance data were collected. The primary performance baseline consisted of course performance data for students taking the two courses from January 1971 to the beginning of the incentive program (September 1971). However, before these data for the 21 months could be used, it was felt necessary to control for other factors. It was possible that students taking the course at different times of the year would show consistent differences in ability. To assess this possibility, the data were broken down by month of graduation. These data for final course grade for the two courses are presented in Fig. 2. The figure indicates that performance was basically constant for the WX course. This finding is further supported by the data in Fig. 3 which show the month
by month mean AOE General scores for the two courses. Here, ability of the WS course was relatively constant. Thus, for the WS course, cyclical trends were in evidence.

However, the picture is quite different for the AER course. Fig. 2 shows that final course grade was generally decreasing over time, grades increased somewhat in August-October of 1970. The ability also does not show as much of a gradual decrease until toward the end of the baseline period, but an increase in ability in August-October 1971 indicated. This pattern is confirmed by Fig. 4 which gives month by month data on time to complete the course. Time to complete increased toward the end of the baseline, but students showed a major increase in speed in August through October.

The decrease in ability and performance toward the end of the baseline supports the information we had been given by the AER supervisor, permitting. They had told us that the quality of the students they were getting had decreased since the beginning of 1971. The increase in performance (speed and exam scores) around August to October is also explainable. The students who graduated from the course in the
September and October given the amount of time they took to complete the course) entered basic training in the beginning of the summer. Thus, most students most likely had finished a full year of school when they entered. Students who entered during the school year are much more likely to be school dropouts or enrollees who left school to enter the Air Force. Thus, even though differences in ability or actual years of education may not differ for the enrollees who enter in early summer, their placement in school is probably higher than for enrollees who enter the course at other times. This greater academic orientation of the students who graduate from the course in September and October could contribute for the greater performance exhibited by the course at these times.

It was created a problem for the determination of the appropriate baseline for AER performance. Recall that the design was such that a baseline was constructed from the year before study was started, and a Hawthorne condition was employed early in the program. Although it was unknown at the time, the Hawthorne condition (from September and October) exactly coincided with the increase in performance brought on by the more highly academically-oriented June graduates from school. Consequently, this group is not a meaningful comparison point for evaluating effects of the incentive systems.

However, the original baseline was still available. But there were problems here as well. First, ability was dropping from the beginning of the year-long baseline to the end, and a continued to drop through the incentive program. Thus, taking the whole year as a baseline would represent an ability level substantially higher than the ability level of students in the course at the time of the incentive program. Consequently, a group was selected as a baseline students who graduated from the course from January 1971 until the start of the first condition (June). This group was still higher in ability than the group in the incentive program (AER: General of 66.7 vs. 5.2 SD = 15.7 and 16.8, respectively) but it was still closer than using the whole baseline.
Another set of baseline data was also collected after the incentive program was completed. It was felt that an indication of the effects of removing the incentive system would be valuable. To obtain these baseline data, performance measures were obtained on a sample of students from each course who started and finished the course after the incentive program was terminated. Instructors had also been asked to continue administering the attitude instruments to the students in this pre-incentive system baseline, but there was apparently some confusion about this, and they did not do so. Consequently, only performance data was available for this pre-baseline.

Finally, another type of baseline data was also collected. It was felt to be possible that conditions at the base could have affected the performance of the students in general. Such things as changes in administrative, political or social incidents, etc., could have some effect on performance. Consequently, performance data from five other courses (Aerospace Ground Equipment Repai ran, Aircraft Environmental Systems Repiaran, Aircraft Structural Repairan, Jet Engine Mechanics, General Purpose Vehicle Repairan) at the base were collected. Samples were taken of 25 students who graduated from each course at each phase of the incentive project (baseline, pre-incentive project, incentive systems 1. 2, and 3, and post-baseline). Thus, six samples of 25 each were taken for each of the five courses. Means for each of the conditions across the five courses were computed. These means were baseline = 83.1, Incentive 1 = 83.9, Incentive 2 = 83.8, Incentive 3 = 83.4 post baseline = 83.1. Thus, the maximum difference was 1 point on a 50-point exam. These data indicate that there were no strong baseline performance changes during the time the incentive program was operational.

Checks on the Manipulations

In order to assess the effectiveness of the manipulations, questionnaires were designed and given to students during the Hawthorne condition and once during the course of each incentive system. The questionnaires were given about half way through each condition. The purpose was to increase (a) perceived relationships between performance and rewards (P-R), and (b) perceived relationships between effort and rewards (E-R). It was expected that P-R perceptions would be higher for the last incentive system than for the Hawthorne condition. Furthermore, P-R perceptions should be lower in the second incentive system than in the first since the second incentive system did not have rewards to performance, but radical effort. Finally, P-R perception should start an increase from the last incentive system to the second since the second system was designed to give rewards on the basis of effort.
two systems were weak, the first system resulted in some time savings (1 day) and the second system, where effort-reward expectancy should have been higher, showed 2.5 times the effect of the first system. Thus, while the strength of the effects is small, the results are in the direction predicted by the theory.

In the third system the results were more impressive, saving 30.5 days. This system used the effort-based rewards and added the financial rewards. Since this implies that the addition of the financial rewards resulted in a total outcome package which was highly valued, it would have been desirable to also have a condition which gave these more valued rewards on the basis of raw performance. Had it been possible to run such a condition, it would have given much clearer information as to the effects of the expectancy.

Overall, then, the results give some support to the expectancy theory model, but the effects are far from overwhelming.

Turning to the applied side of the research, two central questions emerge from these findings. The first deals with the differential effects of the three systems, especially for AER. The answer seems to be fairly straightforward. Only when fairly strong incentives were used, in this case financial incentives, did the program have a substantial impact on performance. Apparently, the incentives used in the first two conditions were not strong enough. This is not to say that nonfinancial incentives would not be effective, and, in fact, had we been able to use the more powerful incentives such as choice assignment, the results of the first two systems would probably have been more impressive.

The second question deals with why the system works better for the AER course than for the WX course. Probably the major reason for this is that, as discussed above, performance of the WX students was so high. Clearly the WX studies did not have much reason to improve. To the extent that these students saw it as very difficult to improve their performance, the strength of the incentive system was weakened.

The question remains, however, of whether or not the incentive systems were cost effective. Although a sophisticated cost-effective analysis was beyond the scope of this project, a crude approximation can be made. The cost of the first two systems would run approximately $300 week for both supplies and 1 man-week required for record keeping. The third incentive system would add $1,500 week for financial incentive. It has been estimated (Air Training Command/Management Analysis Branch, personal communication) that each training day costs the Air Force between $25 and $35/student. The lower estimate is composed of merely the pay, allowances, support costs, and administrative overhead associated with trainees in a holding status, while the higher figure reflects additional costs associated directly with training: instructors, courseware, etc.
The manipulations in the first and third systems, seem to have been successful. Perceptions of performance-reward instrumentation clearly rose in the first session, and it would have been almost impossible for subjects to have perceived that rewards went up to the third system, however, increased manipulation for system 2 did not take. Subjects did not seem to perceive these rewards were related to effort to the second agent than the first. Consequently, we do not have a clear test of the effects of increased expectancy.

It is conceivable that had the second system been longer or the magnitude of effort-reward-expectancy been more sensitive, the second system might have had a better showing. Another explanation is related to the fact that the correlations between derived effort scores and actual performance were, in fact, quite high (median r = .85). Even if rewards were seen as perfectly related to effort, since the "effort"-performance relationships were so high, no difference could probably have been found between effort-reward and performance-reward perceptions. Thus, no differences in effort-reward perceptions from system 1 to system 2 could be expected.

However, the overall pattern of results indicates that the incentives used in the first two systems were too weak to have had much effect. This, even if high effort-reward expectancies have been generated, we could not have expected such effects.

However, before such a pessimistic conclusion is reached, we must explore another finding which bears on this condition. Namely, the positive systems worked better for AER than for WX. Recall that when priority is peripherally not meaningful, performance effects are indicated for WX, however, there is evidence for effects in AER. Specifically, system 1 saved 1 training day, system 2 saved 2.5, and system 3 saved 11.5 days. While these first two effects are small, all three are significant (discussed below) and all were obtained without substantial increases in success rates.

Our explanation of the differences between the two courses rests on the effort-performace expectancy variable. The students in the WX course, I argued out, were very high performers before the system was instituted, in fact, their mean block performance across all students was about 88%. Thus, they were near the upper limit of performance. While attaching effort to outcomes to higher performance substantially increased the value of high performance, they probably saw little chance of increasing their performance. In fact, 88% corresponds to missing just over 2 items on a 10-item test. Thus, the system may not have been effective for WX simply because effort-performance expectancy was too low.

This line of reasoning suggests that the only clean test of the model is for AER course. Here we find that even though the incentives in the first
mechanics of the system. A group of agents dealt with this question by measuring reactions to such things as problems for instructors, restrictions on incentives, amount of paper work involved, delay in getting results, and test points. Mechanics of the program and value of questionnaires required. The sum of these items offered some encouraging results. The students felt the mechanics were adequate, i.e., a mental response (X = 1.2) and the instructions were only slightly below means (X = 2.6). These findings would seem to indicate that the mechanics of the program were not as being particularly burdensome. In summary: the analyses of the treatment evaluation questions revealed the following findings:

1. Students felt very positive about the system.
2. Instructors felt negative, but only slightly negative.
3. Students invariably felt more positive toward the system than did instructors. This was true for all conditions, for all scales.
4. The students equated the three conditions about equally. They were no evidence whatsoever that they preferred the financial incentive system.
5. The students evaluated the financial incentive system as slightly below the other two.
6. There were no differences between evaluations by WX and AER students.
7. AER instructor were more negative than WX instructors.
8. There was great variability concerning the philosophy of using incentives both on the part of the instructors and the students.
9. The interaction were generally neutral about receiving incentives themselves, but there was a great deal of variability.
10. The mechanics of the system were acceptable.

DISCUSSION

This research had essentially a dual purpose. One emphasis of the project asked the practical question of whether such incentive systems would be cost-effective in Air Force technical training. The other emphasis was evaluating expectancy valuation theory in a field experimental situation. On the theoretical side, the research attempted to use valued reward performance system and, in these rewards to effect system with an increase the magnitude of rewards (system index). For an expectancy theory point of view, the first system attempted to increase performance-reward instrumentality and generate new, valid rewards (indications). The second system attempted to increase (relative to the baseline) both distal performance expectancy and performance-reward instrumentality by just rewards to effort. The third system tried to raise the tolerance of reward-
...clearly indicating that the students felt positive toward the program. Another item read "The incentive program is just another form of military discipline." The overall mean for this item was 1.92, indicating fairly strong disagreement. This was one of the lowest means.

There is also insufficient evidence to examine the relationship between items with large variability. The one item which consistently showed high variability dealt with whether students should be given incentives at all. Students generally said yes, while instructors generally said no, but across both groups of students and both groups of instructors, responses to this item showed large variability. Clearly, there is a great deal of disagreement about using incentives at all.

Another item which showed a great deal of variability for the students dealt with the equity of opportunity to obtain incentives. Both groups of students showed great variability in responding to this item. The mean response was near neutral, but clearly some people felt the systems were not fair than did others.

Fears about using negative as well as positive incentives also showed high variability for both students and instructors. As expected, students generally said no while instructors generally said yes. However, the students were not as against the idea as one might expect. The mean across all students was 2.49, just slightly under neutral.

A final individual item of significance reads "An incentive system such as this should also give privileges (incentives) to instructors." The students agreed slightly (mean = 3.32), but across all instructors the mean (2.93) was a neutral response. The Weather instructors were more in favor of it (3.28) while the Electrical instructors were less (2.75). Among the instructors this item had a very large variability.

The final issue on the treatment evaluation questionnaire deals with the
Summary of Attitude Effects

The results of the analyses of the attitude data reveal the following findings:

1. Social desirability response set was equal across all conditions.
2. Attitudes toward re-enlistment increased strongly, especially in the WX course.
3. Attitudes toward the Air Force were slightly, but consistently, better in the incentive program.
4. Students' attitudes toward the course itself did not change.
5. Attitudes about job satisfaction were slightly better during the incentive program.
6. Attitude effects were generally equal for each incentive system. The only exception was re-enlistment attitudes which were more positive in system 3 than in the other two incentive systems.

Attitudes Toward the System

So far, we have discussed data concerning checks on the manipulation, performance data, and attitude data. In addition, it was felt valuable to assess the students' and instructors' direct reactions to the actual incentive systems themselves. To accomplish this, two questionnaires were developed, one for students and one for instructors. These questionnaires dealt with items about the philosophy of giving incentives, perceived effects of the incentive system on performance and attitudes, logistics of the system, the fairness of the systems, and the support given to the system. A total of 31 items were used in the student form, and 16 were used for the instructor form. All items were presented in a five-point Likert format. These questionnaires were given three times, once during each incentive system.

The last item on the questionnaire was: "Overall, I think the program is a good one for the Air Force." Means for this item are presented in Fig. 1. This figure indicates that the students were indeed very positive about the system. The mean response is above 4.0, and since the maximum scale value is 5.0, this mean value is quite high.

Results from this overall item indicate that the instructors were more neutral but not nearly as negative as one might have expected.

Finally, this figure indicates that, if anything, students preferred the financial incentive system to the other systems. Although the differences are very small, clearly, the instructors preferred the third system least.

Two other individual items related to overall evaluations of the incentive system are worth mentioning. One item read: "I would like the incentive system to continue." The mean response across all students and in all the incentive conditions was 4.95. This was the highest mean of any
TABLE I
Summary of Self- and Instructor Effort Ratings

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.E. ratings</td>
<td>5.6</td>
<td>5.6</td>
<td>4.9</td>
<td>5.8</td>
</tr>
<tr>
<td>N</td>
<td>120</td>
<td>40</td>
<td>144</td>
<td>201</td>
</tr>
<tr>
<td>E.I. ratings</td>
<td>4.4</td>
<td>5.7</td>
<td>6.7</td>
<td>5.1</td>
</tr>
<tr>
<td>N</td>
<td>65</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>S.E. ratings</td>
<td>6.3</td>
<td>6.3</td>
<td>0.8</td>
<td>3.3</td>
</tr>
<tr>
<td>N</td>
<td>120</td>
<td>40</td>
<td>144</td>
<td>201</td>
</tr>
<tr>
<td>E.I. ratings</td>
<td>6.6</td>
<td>6.7</td>
<td>6.1</td>
<td>6.3</td>
</tr>
<tr>
<td>N</td>
<td>108</td>
<td>23</td>
<td>220</td>
<td>68</td>
</tr>
</tbody>
</table>

The standard deviation for each cell may be estimated by the square root of the cell.

7. Analysis of performance adjusted for ability differences indicated that no real changes in exam scores in W.X. (b) As expected, a decrease in S.E. exam scores, but differences were small, approximately 3 points out of a 100 point exam. (c) Speed of completion showed stronger effects. Incentive system 1 saved 1 training day; system 2, 2.5 days; and system 2, 10.5 days.

8. Benefits of instruction generally decreased for both courses. Exam failures (AER) were about equal to the baseline, but effects varied across incentive systems.

9. Problem behaviors (W.X) were lower in the incentive systems.

10. Frequency of controlling sessions showed no real change in AER, but decreased for W.X.

Fig. II. Set and instructor ratings of effort (averaged data).
Incentive Motivation Techniques

Recall that the 11 items on this scale dealt with attitudes toward Air Force pay, efficiency of Air Force operations, Air Force concern about individuals, importance of the job the Air Force is

engaged in, extent to which Air Force uses a person's skills, etc. Means for this scale by course and condition were presented earlier in Table 9. This table indicates the same pattern of findings as the re-enrollment data, but the differences are smaller. The WxN students who showed the stronger diets showed an increase of only six-tenths of a scale point from the baseline to Incentive system 3. While this difference is small, it was consistent across the items of the scale. In 22 comparisons made across the 11 items in the two courses of the baseline mean with the incentive system 3 mean, 21 showed a more positive attitude in the incentive condition.

Examining the individual items, those showing the largest increases under the incentive conditions dealt with the efficiency of the Air Force, the concern of the Air Force about the individual, the quality of the Air Force living conditions, and the training pride at being in the Air Force. Taken as a whole, data from the overall Air Force scale shows that students' attitudes toward the Air Force were slightly, but consistently better in the incentive conditions than during the baseline. As with the re-enrollment attitudes, overall Air Force attitudes rose from the first incentive condition to incentive condition 3. However, even in incentive condition 3, attitudes toward the Air Force were only about neutral.

The next attitude measure deals with the course evaluation. This scale contained 27 items dealing with such aspects of the course as instructor presentation, personal characteristics of the classroom setting, training aids, currentness of training material, quality of tests, etc. Since the course system did not change the courses themselves in any way, it was not expected that student course evaluations should change as a result of the incentive systems. However, it was felt possible that some minimal carry-over might occur, so these data were collected. Table 9 presented previously contains the means for this scale by course and condition. As these data clearly indicate, there were virtually no differences across conditions in attitudes toward the two courses. The absolute value of the means corresponds to a point slightly above neutral on the scale.

The last attitude scale to be considered is the Job Satisfaction Scale. The scale dealt with satisfaction with such aspects of the situation as pay, working conditions, feelings of responsibility, use of abilities, etc. Data on the composite of the 22 items, 5-point scale are presented in Table 9. The data indicate a consistent, but very small, increase in attitude during the incentive conditions. AER students were generally more satisfied than WN students.
12. There was a general trend for the secondary performance measures to be influenced positively by the Hawthorne condition, these effects being frequently as strong as the treatment effects.

13. Ratings of effort showed no real effect in AER but increased somewhat in W.K.

Attitudinal Effects

The second major class of dependent variables was the attitude data. While the effects of any incentive system on performance are of major importance, effects of such systems on the attitudes of those in the system are also of significance. Consequently, several types of attitude data were collected. As described previously, the first attitude area was directed towards attitudinal, job satisfaction, course evaluation, overall attitude toward the Air Force, and social desirability. This last was not really an attitude scale per se, but an attempt to ascertain whether students were using a social desirability response set in completing the attitude questionnaires.

These five scales were physically located on three questionnaires. The attitudinal scale and job satisfaction scales were on separate questionnaires, while the latter for the overall Air Force, course evaluation, and social desirability were intermixed in the third questionnaire. All instruments were in the form of five-point Likert scales with the exception of the re-enlistment scale which was a 9-point Likert scale.

This set of questionnaires was administered as the student graduated from the course. They were given during the last week of the course and were administered by the instructors for that section of the course. However, based on the sample sizes for some of the scales in some of the experimental conditions, each of the questionnaires was not routinely administered.

The first scale to be discussed is the social desirability scale. Means for the 30 items of this scale by course and condition are presented in Table 9. This table presents means of all the five attitude scales, as well as sample sizes, the error term (DSSW) for a $3 \times 5$ analysis of variance of these means, and the probabilities associated with this analysis. Note that in attitude data appear for the post baseline. Due to a confusion in instructions, the questionnaires were mistakenly not administered during this condition.

They might argue that if the attitude measures were highly correlated with ability, the decrease in ability across treatments would make the results difficult to interpret. In fact, this was not the case, as the Pearson Kendall $\phi$ indicate these correlations were low. Also, the decrease was not consistent across conditions.
desirability was near equal across conditions. There was a slight graded increase from the beginning of the project to the end, but the differences were very small compared to a scale point. Due to the large sample size, the differences were statistically significant, but again, the absolute differences were too small to be meaningful. This, even if a social desirability set was being used, affected the means of the various conditions equally, and these means are still interpretable.

The second attitude scale in Table 9 is the re-enlistment scale. It included five items dealing with whether the recrue would re-enlist if they were no more likely to re-enlist for a second tour of duty given the choice, whether he would make a career of the Air Force when, in time to retire, and whether he would advise a friend to re-enlist. The means for these items are presented graphically in Fig. 2. The figure clearly shows a large increase in positive attitudes towards re-enlistment in the incentive conditions and a decrease in these positive attitudes as one goes from the first incentive condition to the third. This was true of both courses. The means for the WR course in the third incentive system were based on only seven cases, but the trend was supported by the WR course with a much larger sample size. Furthermore, the pattern for each of the five individual items is identical to the means presented in Fig. 9.

It is important to note that these differences represent substantial differences in response. The means in the baseline condition represent a response of slightly above "probably not," there is a slight chance I might, but I probably wouldn't." The means in the third incentive system are above "I'd say there's a 50-50 chance, I might or I might not." For another way, across both courses, only 9.5% of the recruees in the baseline were above neutral on attitudes toward re-enlistment, while the figure was 41.8% in incentive system 3.

The next attitude scale to be considered is the scale dealing with overall
respects, however, it is surprising that block exam failures were low any of the incentive conditions. Recall these results were based more speed than some in the Electrical course. One would think except the students would take exams earlier than in the baseline and, consequently, far more often. This increase in block exam failure rates did not materialize in the first and third incentive systems, and net effect in the second system was small.

Data on percentage of probation (WX) per block are presented in Fig. 9. In contrast to the block failure data, above data far from the Wolfe probation is presented. As expected probations were rare in blacks (84 out of 2902 blocks). The figure shows a pattern very similar to those figures The incentive condition resulted in a strong (72%) decrease in probation, over the baseline; the first and third incentive systems were about equal in effectiveness, but better than the second system. Hawthorne condition was an effective in the incentive systems.

The final measure in this group of secondary performance measures was the mean number of counseling sessions. These sessions were held in instructor with a student individually, when the instructor felt the student's performance was unsatisfactory.

Data for both courses are presented in Fig. 10. They indicate no differences for the Electrical course until the post baseline when a decrease was observed. Substantial decreases occurred for the Wolfe course, especially for the first incentive systems. Once again, however, the Hawthorne remains close to the level of the incentive systems, as in post baseline.

Results of Efforts

The third class of performance measures consists of self and instructor ratings of the amount of effort expended by the students in technical training. The items comprising this measure were generated after having viewed instructors on what they felt constituted high effort in the courses. Nine items were ultimately used, each that deals with specific aspect.

![Graph](image)

**Fig. 9.** Mean percentage of probation per block. W.C. Number of blocks upon which WX total mean is based: 124 = 100, 11 = 537, 3 = 220, 7 = 107, 1 = 50, PB = 90.
Incentive Motivation Techniques

In an attempt to diagnose the reliability of the instrument, several parallel checks were made. Test-retest reliability estimates were made with the instrument's ratings by having instructors in the Weather course (18 instructors rating 114 students) and Technical course (32 instructors rating 185 students) rate the same students after a 2-4 day time span. Reliability for the WX instructor was .74 and .76 for the AER instructors.

A second type of reliability estimate was obtained by correlating the ratings of two different instructors on the same student. Values obtained were .51 (N = 77) for WX, and .53/N = 122 for AER.

Finally, internal consistency reliabilities were calculated using the Cronbach's alpha. Resulting values were .91 (N = 134) for WX instructors, .87 (N = 145) for AER instructors; .77 (N = 59) for WX students, and .76 (N = 89) for AER students.

In all, the instrument displayed satisfactory reliability. Instructors were fairly consistent across time and at least moderately consistent across students, and the items seem to be measuring the same construct for the instructors. Likewise, the items of the measure seem to be measuring the same construct for the students. However, there is good evidence that somehow the students were not responding to the instrument in the same way as were the instructors. Correlations between

![Graph showing data points and trend line](image-url)
students' self-ratings and instructor ratings (i.e., ratings on the six-item person) were .21 \((N = 340)\) for WX and .26 \((N = 219)\) for AER. On this, the students and instructors did not agree on their perceptions of effort.

Results of the effort ratings are presented in Table 1. The data presented here is a composite based on the sum of the six items of the instrument. Although the difference across conditions is highly significant due to the large sample sizes, the table indicates that the composite effort rating did not vary much across conditions for either course. The students in both courses felt they were exerting more than average effort (mean effort = 5.0), but the incentive systems did not greatly increase or decrease self-perceptions of effort.

Differences were somewhat greater for instructor ratings of effort in the WX course. Ratings of effort were higher under all three incentive systems than in the baseline but showed a decrease from the no incentive condition to the third. In the AER course, there was a difference across the three incentive conditions, and the increase in effort was equal to the baseline. Ratings of effort increased in the AER Hawthorne condition.

The last item on the effort rating questionnaire dealt with overall effort. It is instructive to examine the data for this item. Mean self- and instructor ratings are presented in Fig. 1. Results for AER course were very similar to the finding with the composite effort score. However, in WX, in both self- and instructor ratings effort was seen as being higher in the incentive conditions than in the baseline. There were no differences between the three incentive conditions.

### Summary of Performance Effects

In summary, analyses of the performance variables indicate the following:

1. There were no appreciable cyclical effects for the WX baseline.

2. Cyclical effects were present for the AER course. The result is the elimination of the Hawthorne condition and using only the most recent grades for the baseline.

3. There were no baseline differences in performance across the various conditions.

4. Due to changes in exams and procedures, the post baseline data is uninterpretable, especially for AER.

5. As expected, subjects saw the incentive systems as increasing performance-related accomplishments. However, effort-related accomplishments did not increase for the second two incentive systems.

6. Due to large decreases in ability across treatments, the no-performance data were largely uninterpretable.
failures in the Weather course was so very low (4 failures in 1500)
figure indicates that while the first incentive condition substantially
and block failures relative to the baseline, the second increased
spikes clearly indicate is above the baseline, and the third incentive system was equal with
as compared to the baseline. The Huxthorne was better than any of the incentive
aged 5.0 hr per students. The past baseline was once again below the baseline. In one
representing a decay
22.5%.

- pattern emerged for
were about equal, an
worthy that at least
resulted in a drop in
also shows that in the
was terminated in
course R1 decre
plotted by the dif
over students in the

5. Mean percentage of block exam failures: AGR. Number of blocks upon which
since the frequency option is based. B = 491; H1 = 437; 1 = 447; 2 = 467; 3 = 467; 4: 423. 4
stronger at lower cost. In this research, greater attention should have been
given to use incentives such as choice of assignment, promotion, and
extra leave. In addition, negative incentives should be instituted for poor
performance such as mandatory study sessions on weekends and mileage
restitution. Second, incentive systems should only be used in situations
where performance can be reasonably expected to increase. Third, the
frequency of reinforcement might be increased. Points or incentives
should be given daily, or at least weekly. Finally, there must be more
social support for high performance. This would be from other students,
supervisors, etc. Reward for supervisors and group incentives should be
explored.

In summary, our results suggest that such incentive systems can be
effective in increasing motivation, but that much stronger effects may be
realized when the appropriate facilitating conditions are met.

REFERENCES

Campbell, J. C., Danziger, M. D., Loether, E. K., & Weick, K. E. Managerial

Campbell, J. C., & Pritchard, R. D. Motivation theory in industrial and
organizational psychology. In M. D. Dunnette (Ed.), Handbook of industrial


and goal model of network motivation: Some theoretical boundary conditions.

Kremer, J. Behavior therapy. In P. Maag & M. Rosecrance (Eds.), A special

Logsdon, E. L. Incentives and motivational effectiveness: A psychological view. New

Nurmi, H. J., & Biesinger, J. C. Manipulating incentives to enhance school learning. Review


Raths, 1. W., & Loether, E. K. Educational attitudes and performance. Homewood, III.:
Dow Jones-Irwin, 1964.


REFERENCE NOTES

1. Frederick, E. A. (1968) "Psychological measures of student attitudes toward
technical training: Reliability and factorial validity." Air Force Human Resources

<table>
<thead>
<tr>
<th>Scale</th>
<th>P</th>
<th>R</th>
<th>T</th>
<th>F</th>
<th>N</th>
<th>Mdn</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social desirability</td>
<td>2.7</td>
<td>2.8</td>
<td>2.9</td>
<td>2.9</td>
<td>3.0</td>
<td>2.9</td>
<td>.001</td>
</tr>
<tr>
<td>N = 156</td>
<td>156</td>
<td>101</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggression</td>
<td>3.7</td>
<td>3.4</td>
<td>4.0</td>
<td>5.1</td>
<td>3.8</td>
<td>3.8</td>
<td>.001</td>
</tr>
<tr>
<td>N = 161</td>
<td>161</td>
<td>106</td>
<td>65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moral development</td>
<td>2.4</td>
<td>2.8</td>
<td>2.7</td>
<td>3.0</td>
<td>3.0</td>
<td>2.9</td>
<td>.006</td>
</tr>
<tr>
<td>N = 155</td>
<td>155</td>
<td>101</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social desirability</td>
<td>3.6</td>
<td>3.6</td>
<td>3.7</td>
<td>3.7</td>
<td>3.7</td>
<td>3.7</td>
<td>.005</td>
</tr>
<tr>
<td>N = 157</td>
<td>157</td>
<td>101</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td>.13</td>
</tr>
<tr>
<td>N = 155</td>
<td>155</td>
<td>101</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACR</td>
<td>29.4</td>
<td>29.3</td>
<td>29.5</td>
<td>30.1</td>
<td>29.8</td>
<td>30.1</td>
<td>.05</td>
</tr>
<tr>
<td>N = 146</td>
<td>146</td>
<td>103</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulation</td>
<td>2.8</td>
<td>2.9</td>
<td>3.0</td>
<td>3.9</td>
<td>3.8</td>
<td>3.9</td>
<td>.006</td>
</tr>
<tr>
<td>N = 156</td>
<td>156</td>
<td>101</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moral development</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td>3.0</td>
<td>2.9</td>
<td>2.9</td>
<td>.02</td>
</tr>
<tr>
<td>N = 157</td>
<td>157</td>
<td>101</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction</td>
<td>2.0</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>.13</td>
</tr>
<tr>
<td>N = 157</td>
<td>157</td>
<td>101</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>3.4</td>
<td>3.4</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>.08</td>
</tr>
<tr>
<td>N = 155</td>
<td>155</td>
<td>101</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The standard deviation for each cell may be estimated by the square root of the N.}

Inspection of the means for the social desirability scale for the two courses indicates that students were responding at the neutral point of the scale in both courses and across all conditions. Furthermore, the variability in responses was quite low. Although it is somewhat difficult to interpret the absolute magnitude of these means, it appears that there was an adoption of a mild social desirability response set. Recall that the items in this scale deal with behaviors that are "socially desirable," but which very few people actually do. For example, "You never make a lone trip without checking the safety of your car as required by Air Force regulations." Agreeing with the item, for most people, represents the use of a social desirability response set. Thus, a mean value near neutral would be somewhat high since it indicates that subjects are neither disagreeing nor agreeing with items which under less desirable response set conditions they would disagree with.

This problem is not too serious, however, since the level of social
both courses are presented in Fig. 7. These graphs clearly indicate that, like the Weather course, the baseline RD was 50 hr per student, but the incentive system increased the RD to 85 hr per student, representing a decrease of 55%. In the Electrical course the decrease was 28%.

Within the incentive conditions the same pattern emerged for both courses. The first and third incentive systems were about equal, and both were superior to the second system. It is noteworthy that at least for the Weather course, the Hawthorne condition resulted in just as strong an effect as the incentive systems. The figure also shows that in the WS course, RD increased after the incentive program was terminated to near the level of the baseline. However, in the AER course RD remained below the level of any of the incentive systems after the program was over. However, this decrease is easily explained by the change in criterion, and the special instructions for poorer students in the post-baseline which was discussed above.

Data on block error failures are presented in Fig. 8. This figure presented data for the Electrical course only since the frequency of block
<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Headphone</th>
<th>CRT 1</th>
<th>CRT 2</th>
<th>CRT 3</th>
<th>PB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>SD</td>
<td>X</td>
<td>SD</td>
<td>X</td>
<td>SD</td>
</tr>
<tr>
<td>Adjusted mean</td>
<td>-0.36</td>
<td>0.1</td>
<td>0.17</td>
<td>0.20</td>
<td>0.10</td>
<td>0.49</td>
</tr>
<tr>
<td>Black performance (Weather)</td>
<td>20.5</td>
<td>0.1</td>
<td>44.8</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Adjusted mean</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Black performance (Electrical)</td>
<td>6.67</td>
<td>0.1</td>
<td>44.8</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
ability sheets than usual, and the examiners who were admitted under the program happened to be entering the resident technical training phase at their Air Force career at the time of our project. The program was discontinued, which explains the increase in ability for the post baseline. I suspect that some researchers have a capacity.

The problem, of course, is to somehow take these ability changes into account in interpreting the data. One way to do this is to analyze derived effort scores. Recall that for the second and third incentive systems, pins were given on the basis of going above a target score, which was fixed in predicted performance. This predicted performance was calculated from a regression equation which used ability data as predictors. Thus, in one sense, the derived effort score is a score which has ability partialed out, and analyses using this score are at least in the direction of controlling for ability.

To this end, derived effort scores were calculated for all subjects and analyzed. Table 5 and Fig. 6 present the results of these analyses. For the WX course, the pattern of results remains essentially unchanged. For the AER score data, the pattern changes somewhat in the incentive conditions. 1 and 2 are equal to each other, and the post baseline is about equal to the third incentive system. However, as in the raw data, the highest performance still occurred in the baseline condition.

The derived effort data for AER scores presents a somewhat different picture from the raw data. Subjects in the first incentive condition finished the course about 1 day faster than the baseline; the second condition about 2.5 days faster than the baseline; and the third condition over 28 days faster. As before, the post baseline was faster than all but discussed above, these results are interpretable due to changes in course structure and exam difficulty. As such they are controlled from Fig. 6.

Secondary Performance Measures.

We now turn to those performance measures which were not under the direct influence of the program but which are, nevertheless, indices of the effectiveness of the program. These include data on remedial instruction, direct failure, probation, and instructor counseling.

Remedial instruction is the first variable to be discussed, consisted of a supervised study session over and above normal class time. The instructor was required to attend these sessions when, in the instructor's opinion, his performance was unsatisfactory. Data for remedial instruction (RD) in

One could argue that an ability was changing across conditions, these secondary performance measures would be difficult to interpret. To this end, analyses were conducted using ability measures at covariates, the exams were almost identical to the raw data. Thus, the post data will be presented and discussed.
A problem which makes the interpretation of the data from the actual operating conditions difficult is that the level of ability of the students in the two courses was changing dramatically over the various conditions. Figure 5 indicates the magnitude of these changes in ability. The figure shows the mean of the four AEC scales (General, Administrative, Technical, Electrical) for each condition and each course.

This figure shows that both courses had a near-linear decrease in ability from the beginning of the program to the third treatment; then an increase for the program was over. While the decrease was strong in the WX course (10 percentage point drop), it was overwhelming in the AER course (17 percentage point drop). It should be noted that these decreases were not due in any way to the program per se. They were due solely to the nature of the students who happened to be admitted to the two courses during the incentive program. From our point of view, this was extremely unfortunate since it makes assessment of the effects of the program a more difficult task. What is even worse is that in the post-incentive, which could also have been used for comparison purposes, ability increased. In the WX course it rose to a level equal to the original testing; in the AER, it remained up to the level of incentive system 1.

We, of course, had no knowledge of these ability changes until after the data were analyzed. Upon obtaining these results we made detailed inquiries and found that these ability changes were apparently due to a program the Air Force was conducting in response to the draft situation. They were testing a procedure of admitting enrollees with lower
Analogous data are presented for the AER course in Table 6. Recall that the AER course is self-paced and that both scores on exams and speed of completion are relevant dependent variables. Also note, in contrast to the WJX data, no Hawthorne condition is present due to the baseline considerations discussed previously. The table indicates that for exam scores the incentive conditions resulted in moderate decreases in performance, averaging about 4 percentage points below the baseline. Furthermore, there were no real differences in performance for the no incentive conditions. Finally, performance decreased in the post-baseline conditions.

However, as discussed earlier, the incentive system for the AER course was designed to alter the speed of completion rather than exam scores as given the great weight. To the extent that going through the materials more quickly results in lower exam scores, we would expect the decrease in score if speed was increased.

Table 6 also presents the speed data for the AER course. The table shows the mean number of hours taken to complete each block. Note the mean hours are always equal to the mean number of days to complete the course. This is true since there were six blocks and 6 hr of class time per day. Therefore, mean hours per block times six blocks divided by 6 hr equals number of class days.

The table indicates that days to complete the course increased for the first two incentive systems, students under incentive system 2 took almost 6 days longer to complete the course than students in the baseline. However, incentive system 3 showed a large increase in speed over the baseline and the other two incentive systems. Students in incentive system 3 finished the course 8 days sooner than students in the baseline. However, in the post-baseline students were substantially faster than any other condition, almost 7 days faster than the third incentive system and 13 days faster than the baseline.

The picture that emerges from these data is that the first two incentive systems had a moderate negative effect on exam scores and speed of completion, while the third system increased speed of completion. The post baseline continues the interpretation substantially since in this rate period exam scores decreased while speed increased.

However, there is information that explains the results in this AER post baseline. First, the criterion for passing the block exams was ultimately lowered during the post baseline period. Thus, one would expect lower scores and faster times. Second, a program was instituted at the end of the post baseline period wherein almost all students were all put on some sort of new and special attention. These two facts taken together indicate that the AER post baseline data cannot be compared with the data in the incentive systems.
<table>
<thead>
<tr>
<th>Values</th>
<th>Raw Count</th>
<th>Pattern</th>
<th>Net</th>
<th>Float</th>
<th>TEFL</th>
<th>T/E</th>
<th>E/F</th>
<th>T/F</th>
<th>F/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
</tbody>
</table>
system was perceived as intended: trainees felt rewards (incentives) were being given on the basis of their performance. However, trainees did not feel that rewards were given on the basis of effort in the second-incentive system.

**Inter-correlations of Dependent Variables**

Table 4 and 5 present the intercorrelations between the dependent variables for the WX and AER courses, respectively. The exact nature of these variables will be made more clear later in the text. The correlation matrix indicates that, in general, there is a substantial amount of unique variance in the dependent variables. The only major exception is the near perfect correlation (0.99) between derived effort for exams score and actual exam score in the AER course. Clearly, these two measures are tapping the same aspect of performance.

**Performance Effects**

The dependent variable of central significance in this experiment, program was performance. Consequently, we shall discuss the performance data in some detail. Three classes of performance data were employed in the study. The first deal with termination performance: the courses per se; specifically, scores on exams in the WX course, scores on exams and speed of completing the course in AER. The second, pump-up performance measures are those behavioral measures which were not directly rewarded by the incentive systems but which still reflect student performance. These include amount of remedial instruction, frequency of blocked failures, frequency of prohibitions, and frequency of instructor counseling. The final performance measure are those dealing with self and instructor ratings of effort.

**Primary Performance Measures**

Performance data for the WX course are presented in Table 6. The table presents mean block performance, sample sizes, and standard deviations.

Inspection of Table 6 indicates that the incentive program did not have a strong effect on performance in the WX course. The largest positive effect was for the first incentive system and was less than 2 points above the baseline. While this increase is highly statistically significant due to the large sample sizes, for practical purposes it is quite small. Furthermore, these data indicate that the third incentive system actually had a very slight negative effect on performance. Finally, when the incentive program was over, performance increased to a level slightly above the original baseline. Once again, however, the differences are very small. Clearly, these findings indicate that the incentive system is not having either positive or negative effect on course performance in the WX course.
To measure R-R perceptions, subjects were asked to rate the changes in their high performance from 25% would result in a high level of satisfaction. In the weather course performance was defined as grades on exams, while performance for the Electrical course was defined as speed and score on exams. Separate items were used for speed and exam scores. R-R perceptions were measured in an analogous fashion.

The means of these questionnaire items by course and treatment are presented in Table 3. The data clearly indicate that R-R perceptions were different from the baseline condition to the first incentive system. The mean baseline R-R perception was 2.38 while the mean for incentive system 1 was 2.93. To compare this difference, tests of significance on the data sets of means making up these values (AER performance, AER speed, and WIK performance) were computed. All were significant, with p < .01 and p < .001, respectively.

The data also indicate, as expected, R-R perceptions decreased on the first incentive system by the second. For the three variables making up the mean in the figure (AER performance, AER speed, and WIK performance) the decrease from the first incentive condition to the second was significant (p < .05) for only AER performance, but each was in the predicted direction.

The data are not encouraging for R-R perceptions. Recall that it was expected that R-R perceptions should increase from the first incentive system to the second. The data indicate this was not the case. There were no differences in R-R perceptions between the first and second incentive systems, this indicates that the second incentive system did not result in changes in behavior toward standards based on effort.

In summary, the manipulation checked indicated that the first incentive system...
In the AER and WNS courses, a very conservative estimate of 1 day's
savings is $40. There was no performance effect in WNS, and in AER
there was a time savings of 1 day, 2.5 days, and 10.5 days for the three
incentive systems, respectively. Since the AER course graduated approxi-
mately 15 students per week, the weekly savings in system 1 was $900 ($40
× 15); for system 2, $1,500 ($50 × 30); and for system 3, $3,000 ($100
× 30). Since the weekly costs of the three systems were $300, $300, and
$600, this resulted in a weekly profit of $300, $200, and $4,400 for three
systems, respectively. This analysis clearly shows that all the systems
would be cost effective, particularly the last two.

Admittedly, these profit figures are rough, and probably reflect certain
unavoidable and generous assumptions. They do, however, give the positive
effect of the cost-benefit ratio of incentives at different levels. The
findings in this study are consistent with the findings in
the literature on incentives and motivation.

The final consideration is how such a system could be improved. Clearly,
the relatively weak effects produced by the present research are in
conflict with the pattern of positive results generally reported for
incentive motivation techniques. Based on review of the literature we
have elsewhere argued (Prindah, De Los, & Van Bergen, 1974) that there
are a series of conditions which must be met to achieve a successful
incentive motivation program. We shall discuss each of these and
comment as to how well our program satisfied them.

1. Incentives must be carefully selected and identified as highly
attractive. This condition was met by the research here. Every attempt
was made to isolate incentives and determine their attractiveness.

2. Since there are individual differences in the attractiveness of
different incentives, offering people in the system to choose their own
incentives from a variety of incentives should increase the attractiveness
of the incentive package. This condition was met.

3. The greater the proportion of all positive outcomes in the person's
environment that are subsumed under the incentive system, the more
powerful it will be. The incentive systems did not satisfy this condition.
The scores controlled by the system were relatively weak compared to
the overall rewards for the training. The third system was better than the
first two, but overall, the incentives were not strong.

4. The incentives must be applied consistently by the system. This
condition was met.

5. The person in the system must perceive that the incentives are
applied consistently. At the start of the system, a problem with this
condition arose in that two students were denied requests for 3-day passes
by a course supervisor. While the reasons for the changes were reasonable, the rumor spread through the students that incentives might not be delivered even if earned. This rumor was stopped and cleaner guidelines for incentive delivery were given to both instructors and students. Apparently, no problems arose after this. Consequently, this condition was met, at least after the start of the program.

6. The rules for winning the incentives must be completely clear, those administering the system and those actually in the system. Except as noted directly above, this condition was met.

7. It may be possible for of people in the incentive system to actually do the behaviors required. This condition was probably met for AER, but not met for WX. As discussed above, WX students had very little room to improve their performance.

8. The people in the system must perceive that it is possible to do the behaviors required. AER yes, WX no.

9. The people in the system must perceive that variations in controllable aspects of their behavior will result in variations in their level of performance and, ultimately, their rewards. AER yes, WX no.

10. The behaviors required must be clearly specified as well as easily measured. This condition was met.

11. Once the system is operational, great caution should be exercised in increasing the magnitude of the incentives for a given behavior. The point system was changed at the beginning of program, but students seemed to accept the rationale for the change, and no student to our knowledge ever mentioned the possibility that high performance would result in a lowering of points awarded.

12. The more frequent the reinforcement (i.e., the shorter the rate of performance to be reinforced) the stronger the system. Due to the structure of the courses, reinforcement occurred infrequently for both courses. WX averaged 1.3 reinforced exams per week. AER averaged 2 reinforced exam per 2.5 weeks. More frequent reinforcement may have been more effective.

13. The behaviors to be reinforced by the system should also be reinforced by any significant others in the person's environment. This condition was not met. The instructors in the squashing area (not the course instructors) did not reinforce; encourage high performance in the courses and generally downsloped the incentive system. Also, the norm of the students seemed to be to avoid extra study and extra effort in the courses.

The preceding discussion suggests several reasons both why the present system did not produce strong effects and how an incentive system could be improved. First, the incentives should be stronger. They must also be

Received March 20, 1975