1977

Intrinsic Interest And Extrinsic Reinforcement: Testing The Overjustification Hypothesis In Ongoing Reinforcement Systems

Paul Richard Davidson

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

INTRINSIC INTEREST AND EXTRINSIC REINFORCEMENT: TESTING THE OVERJUSTIFICATION HYPOTHESIS IN ONGOING REINFORCEMENT SYSTEMS

by

Paul Richard Davidson

Department of Psychology

Submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Faculty of Graduate Studies

The University of Western Ontario

London Ontario

October 1976

ABSTRACT

The Overjustification Hypothesis predicts that token reinforcement programs may have deleterious effects on subjects' intrinsic interest in reinforced activities. Support for this hypothesis comes primarily from studies where a short interval of responding by subjects leads to a single reward, and where intrinsic interest is measured in terms of free play choice of the activity. These reward procedures are fundamentally different from applied token reinforcement programs where multiple responses lead to multiple rewards over many days. The present studies were designed to test the Overjustification Hypothesis under conditions more representative of such applied programs.

In Experiment 1 three pre-schoolers were trained to play with two game-like teaching machines, and the effect of reinforcement for play on subsequent free-play responding was assessed. Two baseline conditions without token reinforcement were alternated with two reinforcement conditions where responding to one machine led to tokens while responding to the other had the same consequences as in Baseline (Training segments). At the beginning of each session free play responding to both machines

was assessed with the experimenter absent, and this constituted the primary measure of interest (Testing Segments).

The reinforcement of play with each machine in its turn caused an increase in play with that machine in the corresponding Test segments relative to play with the unreinforced machine. The results were taken as not supporting the Overjustification Hypothesis.

In Experiment 2, three groups totalling 61 undergraduates were used in assessing the effects of reinforcement for an academic task upon subsequent interest in that task. Students were instructed to read passages of course material for homework, with the help of lists of key words identifying important passages on each page. The task was to write and hand in a brief response to some of the key words to indicate that the passage had been understood. In Baseline there was no reinforcement for this task. During Reinforcement points were delivered for each response to a pre-determined maximum. One group got maximal points for 7 responses/session, an increase from the groups median response rate in Baseline (Increase). A second group got maximal points for 4 responses/session, a decrease from its Baseline rate (Decrease). A Control received maximal points for attendance only. After four sessions of the Reinforcement condition, Baseline conditions were resumed. Comparison of differences in the response rate both between baselines and among groups failed to support the Overjustification Hypothesis.

Possible reasons for the failure of these studies to support the Hypothesis are discussed. The results are taken as supporting the contention that there is no evidence of detrimental effects of reinforcement in ongoing token systems.

ACKNOWLEDGEMENT

The completion of a doctoral thesis is a complex operant, involving many chained responses and a large reinforcer delivered after the terminal link only. In links far removed from the reinforcer the probability of response often drops pretty low. I wish to thank Brad Bucher, my advisor and my friend, as well as Peggy Davidson, my patient wife, for providing gentle prompts at these lows. These people provided much support and encouragement at times when it was most needed.

I wish to thank the now defunct Glenmore Day Care Centre and the parents of S_1, S_2, and S_3.

I must acknowledge the Glendale Adult Training Centre, in particular Bob Cole, for much understanding during the preparation of this manuscript and Mary Lou Bousfield for her typing and her good natured comments.

TABLE OF CONTENTS

CERTIFICATE OF EXAMINATION ii

ABSTRACT iii

ACKNOWLEDGEMENT v

TABLE OF CONTENTS vi

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF APPENDICES x

INTRODUCTION 1

Theoretical Background 2

Empirical Background 8

The Overjustification Hypothesis and Token Economies 14

Summary 22

EXPERIMENT 1 23

Subjects 23

Apparatus and Setting 23

Procedure 26

RESULTS 33

DISCUSSION 49

The effects of Reinforcement on Responding 49

The Validity of the Procedure 51

Relevance of Methodology 56

Summary 59

EXPERIMENT 2 61

METHOD 64

Subjects 64

Setting 64
Procedure 65
RESULTS 71
Response Definition 71
Reliability 71
Number of Responses Handed In 72
Proportion of Sessions Where Responses Made . 73
Checks on Procedure 77
DISCUSSION 94
Summary 100
GENERAL DISCUSSION 101
The Present Studies 101
The Logic of Scientific Proof 103
The Previous Research 104
How Overjustification Relates to
Token Economies 105
Alternative Theoretical Stances 106
Conclusions 107
FOOTNOTES 109
APPENDICES 110
REFERENCES 118
VITA 122

LIST OF TABLES

Table	Description	Page
1	Response Preference	37
2	Number of Errors and Differences in Time of Response	46
3	Comparison of Differences Between Mean Number of Responses	76
4	Comparison of Differences Between Proportion of Sessions Responding	80
5	Mean Grades on Tests (Baseline 2)	82
6	Initial Differences: Mean Responses	83
7	Initial Differences: Proportion of Sessions Responding	85
8	Initial Differences: Mean Grades (Baseline 1)	86
9	Practice Effects: Means	88
10	Proportions	89
11	Sensitivity of Measures: Means	91
12	Proportions	93

LIST OF FIGURES

Figure	Description	Page
1	Number of Responses per Session to Each Task	35
2	Mean Number of Responses (Testing)	40
3	Mean Total Responses (Testing)	43
4	Mean Number of Responses/Subject/Session	75
5	Mean Proportion of Sessions Where Subjects Responded	79

LIST OF APPENDICES

Appendix	Description	Page
1	Mean number of responses (Individual Data) - Control Group	110
2	- Increase Group	111
3	- Decrease Group	112
4	Proportion of Sessions Where Subjects Made a Response (Individual Data) - Control Group	113
5	- Increase Group	114
6	- Decrease Group	115
7	Sum of Responses to Both Machines	117

INTRODUCTION

There has been considerable controversy recently about the possibility that token economies and other reinforcement procedures may have detrimental effects upon interest in the reinforced activities when reinforcers cease to be available. Interest in activities is typically operationalized in terms of responses made to the stimuli associated with the activity while the subject is in a free play situation. In a free play situation other activities are usually available, and arbitrary external reinforcers for responding are absent (Lepper, Greene and Nisbett, 1973). Authors approaching the issue from the viewpoint of attribution theory (Bem, 1965, 1972) assert that there is a destructive interaction between extrinsic reinforcement and intrinsic interest, with a net result of decreased levels of interest (e.g. Lepper and Greene, 1974; Greene and Lepper, 1974b; Levine and Fasnacht, 1974; Notz, 1975; and Kruglanski, 1975). Authors with a more behavioural stance, conversely, assert that reward does not harm interest and in fact may increase it (e.g. Reiss & Sushinsky, 1975b, 1976; Feingold and Mahoney, 1975; Borstein and Hamilton, 1975; and Hoppe, 1975). Both positions have empirical as well as theoretical support and it is the purpose of this introduction to review that support. The focus of the review is primarily to determine what evidence exists to indicate that applied reinforcement systems may have harmful effects on interest.

Theoretical Background

1. Attribution Theorists: The basic prediction of attribution theorists discussing the effect of reward on interest is that under certain circumstances intrinsic interest may be diminished following the application of extrinsic reward. An important statement of this hypothesis was made by DeCharms (1968). The hypothesis was stated in terms of the individual's perceptions of his own motivations and defines intrinsic and extrinsic motivation thus:

> . . . whenever a person experiences himself to be the locus of causality for his own behavior (to be an Origin), he will consider himself to be intrinsically motivated. Conversely, when a person perceives the locus of causality for his behavior to be external to himself (that he is a Pawn), he will consider himself to be extrinsically motivated. (p. 328)

If an individual agrees to engage in an intrinsically interesting response in order to gain reward, he may come to perceive the response as being controlled by external forces, and thus as an activity it becomes less intrinsically interesting. DeCharms predicts that if an individual is engaging in an activity without external reward, and if an external reward is then applied contingent upon that behaviour, the rate or quality of the response will deteriorate after the reward is again withdrawn (p. 329). According to this prediction it is necessary that the reward be salient, that the subject be aware of the contingencies, and that the reward be sufficient to explain the response. While this hypothesis is explicit about the effects of reward, its major constructs (external reward and intrinsic interest) are defined in terms of the perceptions of the subject, without operational definitions

of the terms being provided.

DeCharms' phenomenological theory forms the basis of much of the research done in the area, but empirical findings have called for some revision of the original statement. Greene and Lepper (1975) have restated the theory in terms of information processing. They assert that the administration of reward tends to increase the expected extrinsic value of the task.

> At the same time however . . . engagement in a task in order to obtain an extrinsic reward may simultaneously decrease the expected intrinsic incentive value of task engagement, if the extrinsic reward employed is sufficiently salient to produce self perception of extrinsic motivation. Before this hypothetical "salience point" is reached . . . , the addition of extrinsic rewards will have no effect on task-intrinsic propositions. After this salience point, however . . . increases in extrinsic incentive are assumed to lead to decrements in the expected intrinsic incentive value of the task engagement (p. 5-6).

This revised view, with an increased emphasis on salience of reward, is similar to the theoretical stance of other attribution theorists studying the intrinsic - extrinsic motivation question (e.g. Kruglanski, 1975; Ross, 1975).

2. Behaviour Analysts: The basic theory that interest can be harmed by reward has been extended to the use of applied reinforcement systems. Levine and Fasnacht (1974) have cautioned against the "promiscuous" use of token reinforcement, and have offered advice to prac-

-titioners and teachers as to how to avoid its use. In the same vein, Greene and Lepper (1974b) have asserted that the use of reinforcement can "turn play into work". Both of these warnings arise out of the work of attribution theorists.

Behaviour analysts have difficulty responding to such warnings. One basic assumption of operant behaviour analysis is that behaviour is determined by its consequences. Thus the concept of intrinsic interest, a variable with no observable properties by which its value can be determined, is difficult to reconcile with operant empiricism. Skinner (1953) has lumped interest with several other "explanatory fictions" and contends that"one who readily engages in a given activity is not showing interest, he is showing the effect of reinforcement." (p. 72)

Its lack of objective definition has led the term interest to be largely ignored by behaviour analysts. However, recently, Lepper and his associates have operationalized interest in terms of choice of activities in a free play situation (e.g. Lepper et al., 1973; Greene and Lepper, 1974a). Given such a situation where a reinforcer has been applied to a response and then withdrawn, there is no empirical basis in the operant literature for predicting a depression of choice of that response compared to baseline. On the contrary, it is possible that the pairing of the response with reinforcement might be predicted to increase the probability of the response, by giving aspects of the task properties of a secondary reinforcer.

Thus behaviour analysis does not predict the detrimental effects of reinforcement that attribution theory does. Despite this it has been suggested that operant and attribution approaches be combined,

(Lepper and Greene, 1976) and Kopel and Arkowitz (1975) present a comprehensive review of the theoretical and empirical bases for such a marriage. There are, however, problems associated with combining the two literatures, and these will be discussed below.

3. Complications in the Theory: Any theoretical statement on the effects of reward on interest is complicated by difficultites in distinguishing between intrinsic and extrinsic motivation. It is not practicable to state that intrinsic motivation requires a lack of external consequences since almost any response an individual may make will have some consequence in the environment. DeCharms attempted to overcome this difficulty by distinguishing between the terms on the basis of the subject's perceived locus of control. Thus, regardless of environmental events, if a subject viewed his response as being controlled by objective rewards or results external to himself, that response is, by DeCharms' definition, extrinsically motivated. Likewise, if the subject views his response as being under his own control, then no matter what environmental consequences the event may lead to, that event is by definition intrinsically motivated.

Unfortunately, this definition leads to practical problems when one comes to discriminate between responses under control of external rewards and those under intrinsic control. For example, authors in the field view the response of making lines on paper with felt pens by children in a free play situation, as being intrinsically interesting (Lepper et al., 1973; 1974a; Ross, Karinol and Rothstein, 1976). But it is possible to construe this activity as being motivated externally by its results, the lines on the paper. Let us suppose that we selected a group of children from a nursery school, who, in a 60 minute free

play situation played with dried out felt markers which made only faint lines. Let us further suppose that subsequently the group is split, one half being exposed to playing with new felt pens (i.e. with a good supply of ink), the other half being exposed to the same old pens, attached to apparatus dispensing candy contingent upon a line-drawing response. Let us suppose that after the period of free play under the new conditions, both groups were to show an equal increase in the length of lines drawn. Given such an increase, the new feedback (candy or bolder lines) constitutes a reinforcer for each group either under an operant definition, or in terms of DeCharms' definition - objective results or reward external to the subject.

Which of these groups would be expected to show the destructive effects of external reward on intrinsic interest? The group which received new pens is identical to the group which Lepper et al. (1973) called intrinsically motivated; but is the candy group motivated intrinsically or externally? According to DeCharms both groups will be intrinsically motivated, unless they should come to perceive their actions as motivated by external events. It is unfortunate that nowhere does he provide a description of how this perception on the part of the child may be discovered by the experimenter.

Since no clear objective distinction may be made between intrinsically and extrinsically motivated behaviours, how can Lepper and his associates (1973) know that subjects are making the distinction subjectively? If there exists a crucial difference between black marks on paper and tokens, which could allow a response leading to one to be termed intrinsically motivated, and a response leading to the other to be termed extrinsically motivated, this crucial difference is

not an observable or environmentally available feature of the stimuli.

A complication arising from this lack of objective distinction between extrinsic and intrinsic motivation is illustrated by a study by Greene (1974) in which, when one group failed to show the expected decrease in interest following reward, it was concluded that the added external reward was not salient for that group, and for this reason had failed to interfere with the group's high level of intrinsic interest. Thus an external event which did not function in a fashion predicted by the hypothesis was then labelled as not being perceived as an external reward. Such an interpretation makes it possible to refute any study not supporting the hypothesis by simply stating that the reward was not perceived as such (Reiss and Sushinsky, 1975a, 1976).

The fact that the terms intrinsic and extrinsic motivation are defined entirely in terms of events internal to the individual greatly complicates research in the area. For example:

> The most serious problem is that the phenomenon is merely named, not explained. Labelling a behavior as intrinsically motivated begs the question of the theoretical nature of the process through which the behavior has come to be a motive. (Calder and Staw, 1975b, p. 599)

Despite these complications, a great deal of research has been done investigating the relationship between intrinsic and extrinsic motivation, and the issue has received more than passing attention from both behaviour analysts and attribution theorists.

Empirical Background

Much research has been done investigating the effect of reward on interest, using various operational definitions of reward and of interest. Before discusssing the various studies it would be best to define the terms to be used in the discussion.

a) Overjustification Hypothesis: This is the proposition that a person's intrinsic interest in an activity may be reduced by inducing him to engage in that activity as an explicit means to an extrinsic goal (Lepper et al., 1973).

b) Overjustification Effect: This will refer to the above reduction in intrinsic interest as a result of reward.

c) Interest: Since various studies use different operational definitions of interest, each study will be discussed in terms of the specific measure of interest used. The general term interest will refer to the frequency of response or the proportion of time an activity is chosen or performed by subjects in a specific situation.

d) Reward: As with interest, the term reward varies with the studies being discussed, and thus the rewarding stimulus will be specified for each study. The general term reward will refer to tangible objects given to subjects by the experimenter contingent upon a certain response, where the delivery of the object to that subject has not been shown to be reinforcement (see below).

e) Reinforcement: The term reinforcement will be reserved for stimuli which have been demonstrated to increase the rate of a response upon which they are contingent.

1. Early Investigations of the Effect of Reward: Early work in this field operationally defined interest in terms of measures of

selective attention and from attitude questionaires. Parker and Nunnally (1966), for example, showed that the association of neutral nonsense syllables with the delivery of money increased subjects' positive evaluation of, and selective attention toward these stimuli. Walls and Divista (1970) obtained similar positive attitudinal changes using Greek letters as neutral stimuli in the same paradigm. By contrast, Kruglanski, Allon and Lewis (1972) showed negative effects of reward on ratings of task enjoyment, when they gave subjects "prizes" (toys) for engaging in certain games.

A major weakness among these studies lies in the operational definitions of interest which were employed. A validation of the attitudinal measures was not undertaken, and while selective attention measures may have face validity, the relationship between selective attention or attitude questionaires and subsequent behaviour in the situation is far from clear.

2. Work With Adults: Deci (1971, 1972) attempted to investigate the effects of reward on interest, using money delivery as a reward, and defining interest not in terms of verbal evaluation, but in terms of behaviour in the situation. University students were asked to perform "Soma" puzzles in the presence of an experimenter, and were subjected to varying levels of monetary reward. The results of this study, while only marginally significant, were taken as supporting the Overjustification Hypothesis. However, the conclusions have been hotly criticized (Calder and Staw, 1975; Reiss and Sushinsky, 1975). Among the

criticisms is the fact that Deci's studies confound reward conditions with level of performance. Subjects who were rewarded for correct responses may have worked faster or more accurately than did their unrewarded counterparts, and it may have been this difference in performance, and not reward directly, which produced the subsequent decrement in responding. Deci's failure to control for the effects of practice is not unique, and as will be shown below, the error has been repeated in many studies which derive from this work.[1]

3. Work With Children: (a) The Stanford Research: Lepper et al., (1973) attempted to remedy some of the weaknesses of the Deci research, and reported similar findings with children. Subjects were selected from children in a nursery school who initially demonstrated spontaneous use of felt pens in a naturalistic free play situation. They were assigned to one of three groups. In two groups children were asked to engage in the drawing activity, and given a reward (certificate) for their participation. In one of these groups children were told in advance of their performance that they would receive a certificate (Expected) and the other group received certificates without warning, after they had engaged in the activity (Unexpected). Children in the Control goup received no prize for participating. For all groups the reward manipulation was carried out in a special setting with the experimenter present and distributing rewards. After 7 to 14 days the task was again made available in the free play situation, and the amount of time that subjects spent playing with the pens was

recorded. The Expected group spent significantly less time engaging in the drawing activity than did the other two groups (which did not differ from one another), and this result was taken by the authors as supporting the Overjustification Hypothesis.

However, in the study, performance variables were confounded with reward, to the extent that the quality of the pictures drawn by the Expected group was rated as inferior to that of the pictures of the other two groups. Moreover, as Feingold and Mahoney (1975) have pointed out, the observation procedures employed were highly questionable. Observers, although highly reliable (inter-rater correlation = .99) in their ratings of time spent engaged in the activity, were not independent, in that the two raters sat side by side, in the same room, with easy auditory and visual access to one another's mechanical recording devices.

Greene and Lepper (1974a) attempted to replicate the original finding, and to extend it by examining the effect of experimenter demand, using substantially the same design. While the experimenter demand manipulation showed no effect, the Overjustification Effect was replicated showing decreased play with the pens following reward. As in the above study, there was confounding of reward condition and performance. "Differences in the children's performance heralded differences in their subsequent interest" (p. 1143-1144) to the extent that the drawings differed not only in quality, but also in quantity, between the experimental and control groups.

In a second replication (Lepper and Greene, 1975) the degree to which the child thought himself to be supervised was varied. In addition to the expected Overjustification Effect, it was

also shown that the target response decreased as the amount of apparent surveillance increased. Unfortunately, while this study does constitute a replication of the Overjustification Effect, using a different task (playing with puzzles) and a different reward (access to nursery school type toys), it too is weakened by the fact that the reward condition was confounded with performance. In this case the rewarded groups solved the puzzles faster than did the control subjects.

These three studies are taken by the authors as supporting the Overjustification Hypothesis. However, a major difficulty is the alternative explanation that the differences in performance may be accounted for by differences in practice between rewarded and unrewarded subjects. In all three studies there is a covariance of reward and practice which tends to obscure the causal determinants of the Overjustification Effect. Greene and Lepper (1975) have recognized this weakness. They point out that there are several studies where the effect is demonstrated without an accompanying practice effect (e.g. Ross, Karniol and Rothstein, 1976 (reviewed below)). On this basis they argue that differences in practice are not necessary for the effect to be observed, and that while individual studies may have weaknesses, they assert that the literature as a whole supports the Overjustification Hypothesis.

(b) The Waterloo Research: Ross and his associates have conducted studies which tend to support the Overjustification Hypothesis. Ross (1975) replicated the effect using a drum playing task, with candy as a reward. The study also

demonstrated that reward must be salient for maximal Overjustifcation effects to be observed.

In this study, nursery school children were taken to an experimental room and invited to play a drum. Two groups were promised a reward (neither was told what it was) and one additional group neither received nor was promised a reward. The two rewarded groups differed from one another in that one group was informed that the reward was hidden under a conspicuous box in the experimental room (Salient) while the other was not told where the reward was. After five minutes of drum playing, in which the three groups did not differ with respect to number of beats played, subjects were tested to see how much time they played with the drum in a free play situation. This test was carried out immediately and again after five weeks. At both assessments the Salient groups played significantly less than did the other two groups.

Ross et al., (1976) again replicated the Overjustification Effect using a drawing task. Subjects rewarded with candy for engaging in a drawing task tended to spend less time playing with the drawing pens (Bic Bananas) in a subsequent free play situation immediately after treatment and with the experimenter present.

The above studies together with the Stanford research form what is seen by proponents as support for the Overjustification Hypothesis (Greene, 1975). These studies must be seen as analogues of the token reinforcement literature, in that none of the studies took place in an ongoing token system, and in fact none delivered more than one reward per subject. Yet on the basis of the

Overjustification Hypothesis, the effect observed in these studies has been predicted in applied token systems.

The Overjustification Hypothesis and Token Economies

Some authors have asserted that the use of reinforcement procedures can "turn play into work", and that care should be taken in designing token systems so that the adverse effects of reward can be avoided (Levine et al., 1974; Greene et al., 1974b). These assertions derive from the Overjustification Hypothesis and its supporting data, the studies reviewed above. There is no direct evidence that ongoing token systems can be harmful.

Despite the lack of direct evidence, warnings such as these are being heeded, and in a newsletter[2] for a group concerned with the education of exceptional children, it was reported that scientists at Stanford University had found that "giving out gold stars and praise to children who are enjoying what they are doing can turn pleasure into drudgery" and cautions "that a teacher should try to maintain natural interest . . . using external rewards only when necessary."

The question being addressed by this paper is: "what evidence is there that the Overjustification Effect occurs in ongoing reinforcement systems"? There is some research which speaks to this question.

1. Applied Research: The studies described so far do not directly show the applicability of the Overjustification Hypothesis or its supporting data to applied token systems. There are two differences between the conditions described in those studies and conditions in token systems which may be

critical. First, in those studies only one reward was delivered to each subject, while in an ongoing token system, a given response may be reinforced hundreds of times. It has been hypothesized that some feature of this single reward system may be the factor which leads to the Overjustification Effect, in which case the effect might not be observed in an ongoing system (Reiss et al., 1976). Second, the rewards used in those studies were never shown to be reinforcers in the operant sense (to increase the response upon which they are contingent). In light of the nature of some of the rewards used (certificates, an opportunity to play with toys found in any nursery school) it may be questioned whether they were reinforcers.

Lepper and Greene (1976) assert that the second difference is unimportant to their theory since they are investigating the cognitive effects of reward. While such an assertion may be valid, it harkens back to the problems arising out of the cognitive definitions of interest. The use of cognitive definitions of reward make the theory of Overjustification foreign to token systems where reinforcers are not arbitrarily but empirically determined.

Clearly either of these two differences may be sufficient to limit the Overjustification Effect to the laboratory. Applied research is necessary to resolve this problem.

Greene (1974) conducted an applied study which he takes as supporting the Overjustification Hypothesis. The

study was carried out with a grade school population, using four different school activities among which the children were allowed to choose. After a Baseline period the entire class was exposed to a reinforcement procedure. One group was reinforced for performance of its preferred tasks, another for tasks it had not preferred during baseline, and a third group was allowed to choose the tasks for which it was to be reinforced. Each group was compared to the rest of the class for control purposes. Subsequently the class was returned to its normal functioning, without reinforcement. Reinforcement consisted of points, which counted toward grades and which were delivered twice a week.

In all groups there was a decline in interest in the reinforced tasks from the reinforcement phase to the final Baseline, and the overall results seem to indicate support for the Overjustification Hypothesis. However, the following factors make the results difficult to interpret: each task became progressively more difficult as the child made more and more responses to the task. Since the groups which were reinforced for specific tasks tended to engage in these more under conditions of reinforcement, they made more responses than did the appropriate control group, and thus were faced with more difficult and advanced tasks than were control subjects, when the reinforcement was removed. In this study then, the decreased responding effect may not have been due directly to reinforcement, but to differences in fatigue, or difficulty which accompanied this reinforcement.

Unfortunately, this confounding of reinforcement and task difficulty makes the results of this study difficult to interpret.

Contrary to the Overjustification Hypothesis, there is a body of literature which indicates that under some circumstances the use of reinforcement can increase responding to a given task following withdrawal of the reinforcer (i.e. can increase interest).

Betancourt and Zeiler (1971) investigated the effect of reinforcing different tasks normally chosen and carried out by nursery school children. Following a baseline period where all tasks led to equal reinforcement, subjects were differentially reinforced for either the tasks which they had previously preferred, or those which they had not preferred over several sessions. Subsequently both groups were returned to non-differential reinforcement as in baseline. Both groups showed an increase in choice of the reinforced task when it led to differential reinforcement, and a trend back toward baseline responding during non-differential reinforcement. Contrary to the Overjustification Hypothesis, the group which had been reinforced for its preferred task showed a slight increase in choice of that task after training ceased. Unfortunately, the study was not designed to investigate how this increase was caused, and it confounded practice and passage of time with reinforcement.

Bucher and Weisz (1975) demonstrated a similar increase in choice of certain tasks after those tasks had led to

reinforcement. On Training Trails children were presented with two games, one at a time. One game led to reinforcement, the other did not. On intermixed Test Trials subjects were offered a choice between the two games, and were clearly informed that no reinforcement was available on those trials. In an ABA design it was shown that reinforcement of one game on Training Trails led to increased choice of that specific game on Test Trials, regardless of which game was reinforced. The authors concluded that unconstrained choosing between games can be increased by the application of reinforcement. This finding is also contrary to the predictions of the Over-justification Hypothesis under conditions of multiple reinforcement delivery over several sessions. One weakness in the study is the fact that interest was assessed in the same setting, with the experimenter present, very shortly after the reinforcer was delivered (20 seconds). An attempt to evaluate preferences after a longer period consisted of only one trial, and while, in fact, the preferences remained on this trial, several trials over a number of sessions would have revealed more about the long term effects of reinforcement.

Feingold and Mahoney (1975) used a dot-to-dot task to investigate the effect of reinforcement on interest. Using the number of dots joined by children in a session as a measure of interest, this study showed no decrease in this measure following an extended reinforcement program, but it is weakened by the fact that practice during reinforcement was not controlled.

Reiss and Sushinsky (1975a) replicated the Overjustification Effect using a music-listening task, with candy delivered one time as a reward. However, when they shifted to a multiple reinforcement procedure in a second study, where 10 candy reinforcers were delivered for the same task, they failed to replicate their original study, and in fact showed an increase in choice of reinforced tasks after the reinforcement procedure was withdrawn. Unfortunately the studies were conducted separately, and thus do not provide conclusive evidence that single reward is a necessary condition for the Overjustification Effect to be observed.

In summary, applied evidence of the Overjustification Effect in ongoing reinforcement systems is not strong. The one applied study which seems to support the hypothesis has been shown to be explainable in terms of unrelated events, while other applied studies suggest an increase in interest with reward, an effect contrary to the predictions of the Overjustification Hypothesis.

2. The Token Economy Literature: While there has been a great deal of research into token economies, Greene and Lepper (1975) assert that most of these studies are not relevant to the Overjustification Hypothesis. They point out that there are four criteria for a fair test of the Overjustification Hypothesis: (Lepper and Greene, 1976):

1) Subjects must not perceive a continuation of the instrumentality of the task after the reward is withdrawn.

If the subject believes that further responses will lead to reinforcement, his interest will appear to be maintained.

2) The rewards must not convey any information about subjects' ability or competence, since perceptions of competence are equivalent with perceptions of internal locus of control, and thus could augment interest.

3) The rewards must not train the subjects in any generalizable skills, since such training could lead to the competence necessary to experience the intrinsic satisfactions of the activity.

4) The rewards must be salient and powerful. That is, the individual must view his response as being a means to an extrinsic end.

While there are few token economies which would meet all these criteria, there are three additional factors which are seen as making the literature on token systems irrelevent to the question at hand (p. 27 - 28):

1) Subjects in such programs are typically selected on the basis of low pre-treatment levels of responding (low initial interest) and thus a floor effect may hide the Overjustification phenomenon.

2) Many of these studies do not provide appropriate controls to allow for the examination of after effects, unconfounded by other changes in the setting.

3) Generalization of token systems is typically assessed in the presence of potentially controlling extrinsic contingencies. For example, a teacher, having achieved a level

of control viewed as desirable, has a vested interest in maintaining that control when the program is withdrawn, and may substitute natural reinforcers for tokens.

Greene (1974) asserts that the failure of the token economy literature to support the Overjustification Hypothesis does not indicate that the effect does not occur in token systems. On the contrary he asserts that aspects of the designs and measures used in that literature make it impossible to draw any conclusions about Overjustification. It is the prediction of the Overjustification Hypothesis that token economies should decrease interest in the reinforced task. The possibility that subjects may be trained in generalizable skills, or that they may perceive continuing instrumentality could hide the Overjustification Effect, not because it did not occur, but because interest is not assessed under the same conditions following the reinforcement program as it was prior to it. In addition Lepper and Greene (1976) point out that other external contolling factors may come to control the response following the withdrawal of reinforcement, and that in such cases intrinsic interest is not being assessed, and thus the effect may be hidden. They point to what they term the "consistant failure of token programs to produce...some persistence or generalization of effects to situations in which the token program is not available" (p. 28) and suggest that this failure may be due in part to the Overjustification Effect counteracting the benefits of the programs of reinforcement.

However, direct evidence for the hypothesis from an ongoing token system does not exist. Investigation in such a system has been called for (Greene and Lepper, 1974; Feingold and Mahoney, 1975)

and is warranted by the fact that the theory is now seen by reviewers as applying to such systems (Levine and Fasnacht, 1974; Notz, 1975). The studies which follow describe such an investigation.

Summary

In terms of the question posed at the outset, there exists no direct evidence that token systems or other applied reinforcement procedures produce the Overjustification Effect. While there is empirical support for the Overjustification Hypothesis, it comes from studies using only one session and one reward. On the other hand studies employing more than one session and several rewards generally fail to support the hypothesis. The available data tend to suggest to proponents of the Overjustification Hypothesis that the Effect occurs in ongoing token systems, while proponents of reinforcement systems interpret the available data as not relevant to token systems. Direct investigation, in an ongoing token system, of the Overjustification Hypothesis is seen as one way to begin to clarify this issue.

EXPERIMENT I

This study was conducted to investigate the Overjustification Hypothesis within the context of an ongoing reinforcement system. The reinforcement system employed was one which delivered multiple reinforcers contingent upon several responses over a number of sessions. Based on Lepper and Greene (1976), a study using ongoing reinforcement procedures which meets the criteria for a fair test of the Overjustification Hypothesis should show the Overjustification Effect. That is, subjects who receive reinforcement for a response should show a rate of response which is below baseline, following the withdrawal of the reinforcer. The present study is seen as a fair test of that prediction.

Subjects

Four subjects were selected from a population of normal intelligence children at a day care centre, on the basis of willingness and availability to serve in the study. Subject 1 (S_1) and Subject 2 (S_2) were boys aged five years six months and four years four months respectively at the beginning of the study. Subject 3 (S_3) and Subject 4 (S_4) were girls both aged four years ten months. S_4 refused to come to sessions after the third day, and after repeated coaxings to return, was dropped from the study.

Apparatus and Setting

The apparatus consisted of two completely child operated teaching machines designed for this study, and a marble dropping apparatus described elsewhere (Gewirtz and Baer, 1958). One

teaching machine (Clown Task) operated as follows: the child selected a large red card with three boxes drawn on it, and placed it in a slide in the apparatus. The boxes contained a numerical problem in a matching-to-sample format, consisting of a sample and two comparison stimuli (e.g. the numeral three in the sample box, and three butterflies in one comparison box, and two butterflies in the other). When placed in the slide the card operated switches which prepared the apparatus by enabling logic circuits designating which key was correct (S^+). Two keys were adjacent to the card in the slide, and were illuminated and activated when the card was positioned correctly. The task was to press one of the keys. The consequences of an incorrect response (to the non-matching comparison) were the end of the trial signalled by the darkening of the keys. A correct response led to the flashing of lights in the face of a clown (dubbed Ronald MacDonald) painted on the appratus, the ringing of a bell, and the playing of taped ragtime piano music within the machine for two seconds; followed by the end of trial. If the two keys were pressed within 0.5 seconds of one another, the trial was treated as an error regardless of which key was pressed first.

Aspects of the House task teaching machine (Sesame Street) were similar to the Clown Task. Selecting a small yellow card with three stimuli on it, the subject placed it in a multi-coloured machine, shaped like a house. The cards contained alphabetic problems (e.g. Letter "C" with the comparisons of a cow and a bird picture). The task was again to press the

button nearest the correct comparison stimulus, but in this case the feedback consisted of the ringing of a bell and the illumination of lights in various windows in the house. The windows lit in a haphazard order, and when they were lit, subjects could look in and see pictures depicting one of six characters from the Sesame Street television program.

After a response was made to either machine the subject removed the card and placed it in a common slot in a separate box. This activated a switch which signalled both machines to be ready for another trial after an interval of 10 seconds. Failure to insert the card in the slot resulted in the machines remaining inoperative indefinitely. Pressing a button when no card was in the machine had no scheduled consequences.

The marble machine was used to assess reinforcer effectiveness. It consisted of a box, with two holes in it, into which subjects dropped marbles. Marbles activated a counter in the hole, and were then returned to the subject. In any session, one hole was designated S^+, and when the subject dropped a marble into that hole, he was given a token by the experimenter.

Sessions were divided into two parts: A Testing segment followed by a Training segment. The Testing segment involved the subjects playing with the teaching machines by themselves, while the Training segments were for the purpose of experimenter

intervention. Both were carried out in one room of a mobile laboratory which was equipped with a through-the-wall observation system to allow for clandestine recording of subjects' behaviour. The room contained a long table with two machines on it, separated by the common receptacle for used cards. Unused cards were located on a chair at the the other side of the room in two unordered piles.

Procedure

The procedure of this experiment is divided into four phases In Baseline 1 all the subjects were trained in the use of the teaching machines (Clown and House) and Baseline responding to the same machines in free play was observed. In the second phase (Reinforce Clown) reinforcement was delivered contingent upon a response to the Clown task during training. Responding to both machines was again observed during free play each day. The third phase consituted a return to Baseline (Baseline 2), where neither task was instrumental in reinforcement, where reinforcer effectiveness was determined by means of the marble task, and where free play responses to both tasks were recorded. The fourth phase (Reinforce House) involved the reinforcing of responses to the previously unreinforced House Task during training, and the observation of free play responding to both machines during testing. In all Training segments subjects were required by the experimental procedures to make an equal number of responses to both machines. This was to prevent an artifact of differences in number of responses during Training from obscuring the results.

Pretraining: Prior to the experiment, a pretraining session

was conducted in which the child was assessed to determine that he/she could identify all the stimuli to be used in the experiment; trained in the token system used throughout the experiment; and generally became accustomed to the experimenter and the setting. From a group of cards with animal pictures, shapes or colours on them, the experimenter presented one card and asked the subject to select an identical card from another group. He also asked the subject to verbally name all the stimuli on all the cards. For correct responses subjects received tokens (brass medallions) which they placed on a sheet of paper with 12 X's on it. Tokens were delivered on a continuous schedule of reinforcement initially, but this was quickly faded to about VR 10. When all the X's were covered with tokens, the session ended, and subjects were allowed to redeem their tokens for their choice of sweets, cookies, orange juice, trinkets or toys, all of which had various token costs associated with them, and which were displayed prominently in the experimental room during training.

This token system was used throughout the study. In all sessions except Pretraining, a continuous reinforcement schedule was in effect for the target response.

Baseline: (Non-reinforced training on both tasks) In this condition subjects were trained in the operation of both machines without the application of tokens. In any session the subjects were brought to the experimental chamber where the simple task of identifying the numbers, letters or pictures

used in pretraining was carried out, with correct answers resulting in tokens. Twelve tokens were delivered and then exchanged, in order to equate for density of reinforcement throughout the study. After the subject was given his/her prize, he/she was taken to the machines and asked to play with them. The first session consisted of a Training segment only, during which the experimenter described the salient features of the machines and cards to the subjects, as well as demonstrating their use. He then asked the subject to play with them. Each machine was operated by the child a total of 12 times in any given Training segment, trials being in blocks of six on each machine. The machines alternated from block to block, the starting machine being determined by a coin flip. In the first session prompts were given (both verbal and physical) to aid the subject in the operation of the apparatus, but these were discontinued in subsequent sessions. After each trial the Experimenter would respond: "very good" if Subjects made the correct response; and "oops, that's a mistake" if they made an error. Nothing else was said until at least five seconds after the response was made. After the completion of 24 trials, the child returned to the day school in the company of the experimenter. In sessions where no tokens were given for the teaching machine performance, the rewards were removed from the experimental room prior to the Training segment of the session.

Since subjects had not been trained in the operation of the machines, there was no Testing segment in the first

session. Subsequently, free play choice of each task was assessed (Testing) each day prior to the Training segment. After the child was brought to the trailer the experimenter said:

> "Today I want you to play by yourself with the machines for a few minutes while I make the orange juice. After that we can play together."

With that the experimenter retired to an adjoining room and locked the door. From this vantage point he watched the subject for five minutes, and then returned. Each day the subject was asked if he had played with the machines by himself. If he reported that he had not, he was asked why not, and encouraged to do so the following day. If he reported playing with only one machine, he was informed, "You can play with both machines if you want to, you know." A report of playing with both machines was reinforced by praise from the experimenter. Apart from this, no comments were ever made about the subjects' performance in the Testing segment of any sessions. Subjects always answered the questions truthfully.

During the Testing segment none of the rewards were present in the room, but these were brought in by the experimenter, who solicited aid from the subjects. After the Testing segment and the bringing in of the rewards, the Training segment was carried out as described above.

Reinforce Clown: In this condition the only aspect of the procedure which changed was the delivery of tokens during the

Training segment. Instead of being delivered for a task prior to the training with the machines, tokens were delivered contingent upon playing with the Clown. In order that the tokens should not add any feedback about the skill of the subject at the machine, tokens were given for any response to the machines, regardless of whether it was correct or not. At the beginning of sessions subjects were asked if they would like to play with Ronald McDonald for tokens that day. They were told that they would get no tokens for Sesame Street. When they said that they would (no subject ever refused) they were asked which machine they were playing with for tokens, and a correct response was elicited. In reinforced Training sessions, the token was always cued by the experimenter saying "Here is a token." (After a correct response to the reinforced apparatus he said: "Very good, here is a token." After an error he said: "Oops, that's a mistake, here's a token.") Procedures continued as in Baseline for the unreinforced House task. The experimenter would make no other comments for at least five seconds after the response was made. After that time he would respond briefly to questions posed by subjects. Tokens were traded at the end of each session. Each day each subject was asked, at the end of the session, which machine was his favourite (if he didn't respond he was asked which of the machines he would prefer to take home with him), and which machine he was getting tokens for. In both questions, when the

subject named a machine, his response was met with a non-committal "Thank you".

Baseline 2: During this phase, conditions in the Training segment of sessions returned to the non-reinforced training of both tasks as in Baseline 1. However, instead of gaining tokens for the simple identification task, subjects were given tokens in connection with playing with a marble machine, thus allowing an assessment of the tokens' effectiveness as reinforcers. Conditions in the Testing segment of each session were unchanged from preceding conditions.

The marble machine consisted of a box with two holes in it, into which subjects dropped marbles. One hole was selected as S^+ by means of a coin flip, and subjects received a token if they dropped a marble into that hole on any trial. Subjects were instructed: "Today we are going to earn tokens by playing with a new machine. When I say so, drop a marble into one of the holes. Sometimes I will give you a token." Each trial was then cued: "Drop a marble into one of the holes." If the hole selected was S^+ the experimenter said: "Here is a token", and then after presenting the token went on to the next trial. If subjects selected the other hole, the experimenter made no comment and went on to the next trial. This procedure was continued until twelve tokens had been delivered. It was predicted that if tokens functioned as reinforcers, most responses occurring after the first reinforced trial in any session would be to S^+.

Reinforce House: This condition mirrored the previous reinforcement condition, with the single exception that tokens were

provided for playing with the House instead of the Clown. Use of the marble machine was discontinued for these sessions. In the final session, after Testing was completed, reinforcer effectiveness was again assessed using that machine.

Experimental Design: The present experiment is neither a classical reversal design nor a classical multiple baseline design, but incorporates aspects of both. Since there are two tasks available in each of the daily Testing segments, and since only one of these (at most) is reinforced in any given Training segment, responding during Testing segments to the unreinforced task serves as a Baseline with which responding to the reinforced task may be compared. The two tasks permit a session-by-session evaluation of the effects of reinforcement in Training segments upon interest in the reinforced task as measured in Testing segments. In addition, the Baseline 1 condition compared to Reinforce Clown permits an assessment of the effect on both tasks of the onset of reinforcement for responses to the Clown Task. The Baseline 2 condition represents a significant departure from the multiple baseline design, and was necessary to clarify inferences about causality since the change from Reinforce Clown to Reinforce House involved two manipulations. In this change reinforcement was withdrawn from the Clown Task and reinforcement was provided for the House task. Baseline 2 allows the effects of these manipulations to be separated.

The Overjustification Hypothesis predicts that interest in the task reinforced in any Training segment should decrease in the corresponding Test segment, while interest in the unreinforced task should be unaffected.

RESULTS

The present study was an investigation of the effect of reinforcement for responses to a task upon subsequent interest in that task. The measure of interest employed was the number of responses subjects made to each of the two tasks (House and Clown) while alone with both machines at the beginning of the session each day (Testing Segment). The number of responses to each task was recorded by the experimenter in the adjacent room, and by electronic counters attached to each machine. The two sets of data thus obtained were compared at the end of each session and were invariably in agreement. Data are also reported with respect to the behaviour of subjects in Training segments when the experimenter was present.

Testing Segments: Each day subjects had an opportunity to make a number of unconstrained choices between the two tasks for a period of five minutes. During this period the experimenter was absent from the room, and tangible reinforcement was neither present nor available. The number of responses subjects made to each task during this period was taken as reflecting the child's interest in each task.

The number of responses made by each subject to each task during each of the 17 testing segments is shown in figure 1. The Over-justification Hypothesis makes clear predictions about these data.

- - - - - - - - - - - - - - - -

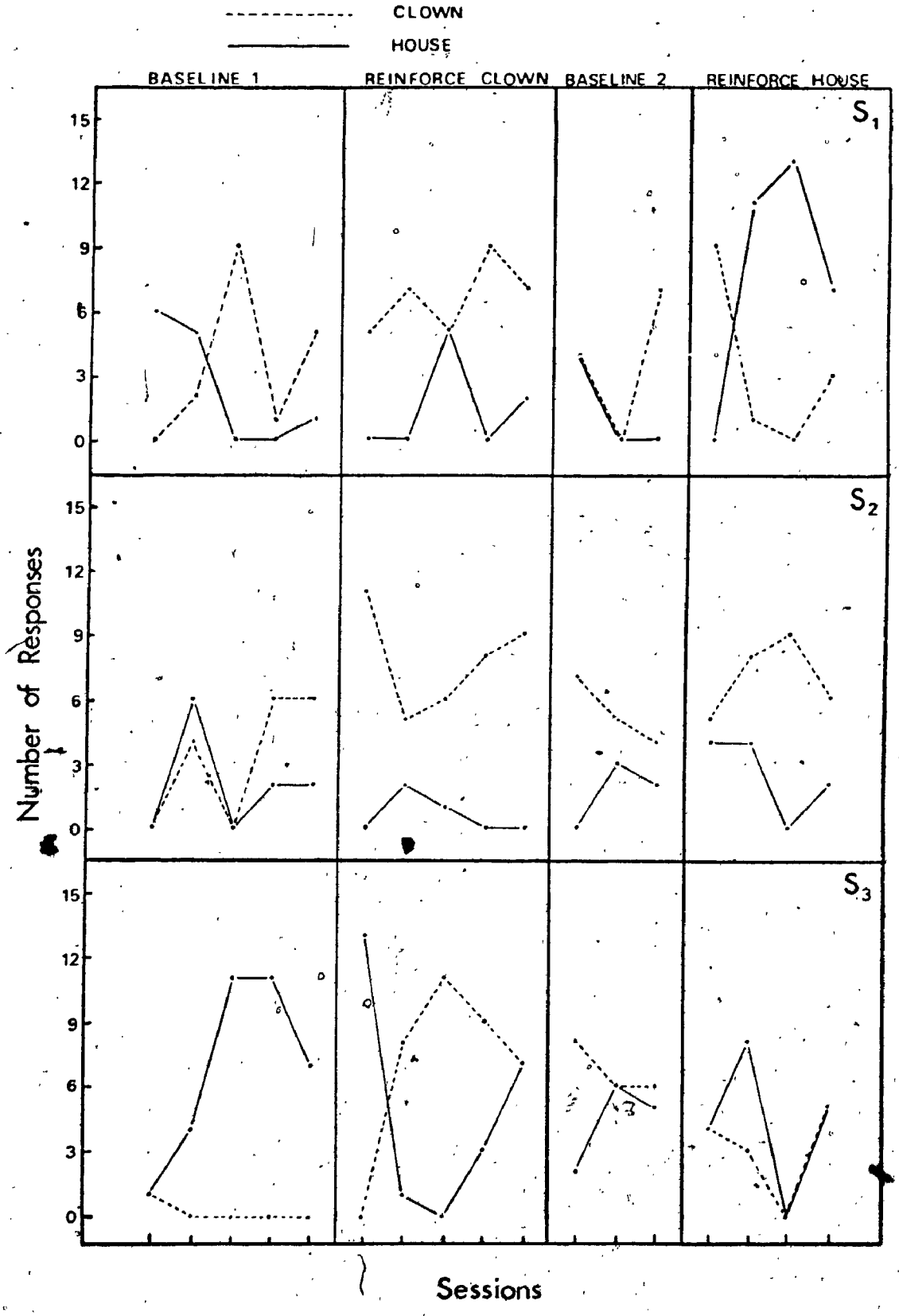
Insert figure 1 about here

- - - - - - - - - - - - - - - -

Since reinforcement is hypothesized to decrease interest, the onset

Figure 1: Number of Responses per Session to Each Task.

Shows number of responses made by each subject to each task during free-play Testing Segments.


CLOWN
HOUSE
BASELINE 1
REINFORCE CLOWN
BASELINE 2
REINFORCE HOUSE
S1
S2
S3
Number of Responses
Sessions
0
3
6
9
12
15

of reinforcement in Training segments should be accompanied by a decrease in choice of the reinforced activity during the corresponding Testing segments. Thus in Testing segments during Reinforce Clown subjects would be predicted to show less interest in the clown task than they did in the Baseline condition preceding it. (Likewise for the house task during Reinforce House)

Inspection of figure 1 yields no support for this prediction during either Reinforce condition. No subject showed a marked decrease in choice of the clown task compared to Baseline 1, when the clown task was reinforced in Training segments, and no decrease in choice of the House task is evident from Baseline 2 to Reinforce House.

In addition to comparisons between Baseline and Reinforce for each task, the present design allows for comparison of changes in responding to the two tasks from Baseline to the Reinforce conditions. Since there were only two tasks available to subjects in each Testing segment, and since only one of these was reinforced in any given Training segment, choices of the non-reinforced task in Testing segments during Reinforcement served as a baseline with which changes in choice of the reinforced task could be compared. This was accomplished by dividing the number of responses to the Clown task by the sum of responses to the Clown and House task for each session (C/(C+H)), and averaging the values obtained within each condition for each subject. Thus relative preference for the Clown task was determined in each condition for each subject. These data are presented in table 1.

- - - - - - - - - - - - - -

Insert table 1 about here

- - - - - - - - - - - - - -

SUBJECT	AVERAGE RESPONSES PER SESSION TO:	BASELINE 1		REINFORCE CLOWN		BASELINE 2		REINFORCE HOUSE	
		number	$\frac{C}{C+H}$	number	$\frac{C}{C+H}$	number	$\frac{C}{C+H}$	number	$\frac{C}{C+H}$
S_1	CLOWN (C)	3.4	0.6	6.6	0.8	3.7	0.7	3.3	0.3
	HOUSE (H)	2.4		1.4		1.3		7.5	
S_2	CLOWN (C)	3.1	0.6	7.8	0.9	5.3	0.8	7.0	0.7
	HOUSE (H)	2.0		0.6		1.7		2.5	
S_3	CLOWN (C)	0.2	0.03	7.0	0.6	6.7	0.6	3.8	0.5
	HOUSE (H)	6.8		5.0		4.3		4.3	
MEAN RESPONSE PREFERENCE			0.4		0.8		0.7		0.5

TABLE 1: Response Preference: Shows average number of responses per session by condition for each subject and ratio of average responses to Clown (C) to sum of average responses to Clown and House (C+H) for each condition for each subject. Mean response preference is the average of the 3 values for (C/(C+H) in each condition. Values greater than 0.5 indicate relative preference for the Clown, less than 0.5 - for House.

A value for (C/(C+H)) greater than 0.5 indicates that relatively more responses were made to the Clown task than to the House task, while a value less than 0.5 indicates the converse.

The Overjustification Hypothesis would predict that relative preference for the Clown task (C/(C+H)) should decrease during Reinforce Clown as interest in the Clown task decreases, and that it should increase during Reinforce House, as interest in the House task (H) decreases. However the data do not support these predictions, and in fact for all subjects the value of C/C+H increased during Reinforce Clown (compared to Baseline 1) and decreased during Reinforce House (compared to Baseline 2). Together the Reinforce Clown and Reinforce House conditions constitute two replications for each subject of increased relative preference for the the task being reinforced during Training segments.

Figure 2 is a graphic representation of responding for the

Insert figure 2 about here

---

three subjects averaged together, by summing daily responses for each task across subjects, and dividing the value obtained by the number of subjects. The results are similar to all the individual plots in figure 1, and it is clear that when the Clown task was reinforced, responding to that task increased relative to responding in Baseline 1 and relative to corresponding responses to the House task (see table 1). In Reinforce House responding to the House task, initially lower than responding to the Clown task, became greater than responding to the Clown task. It increased compared to Baseline 2 House task responding as well.

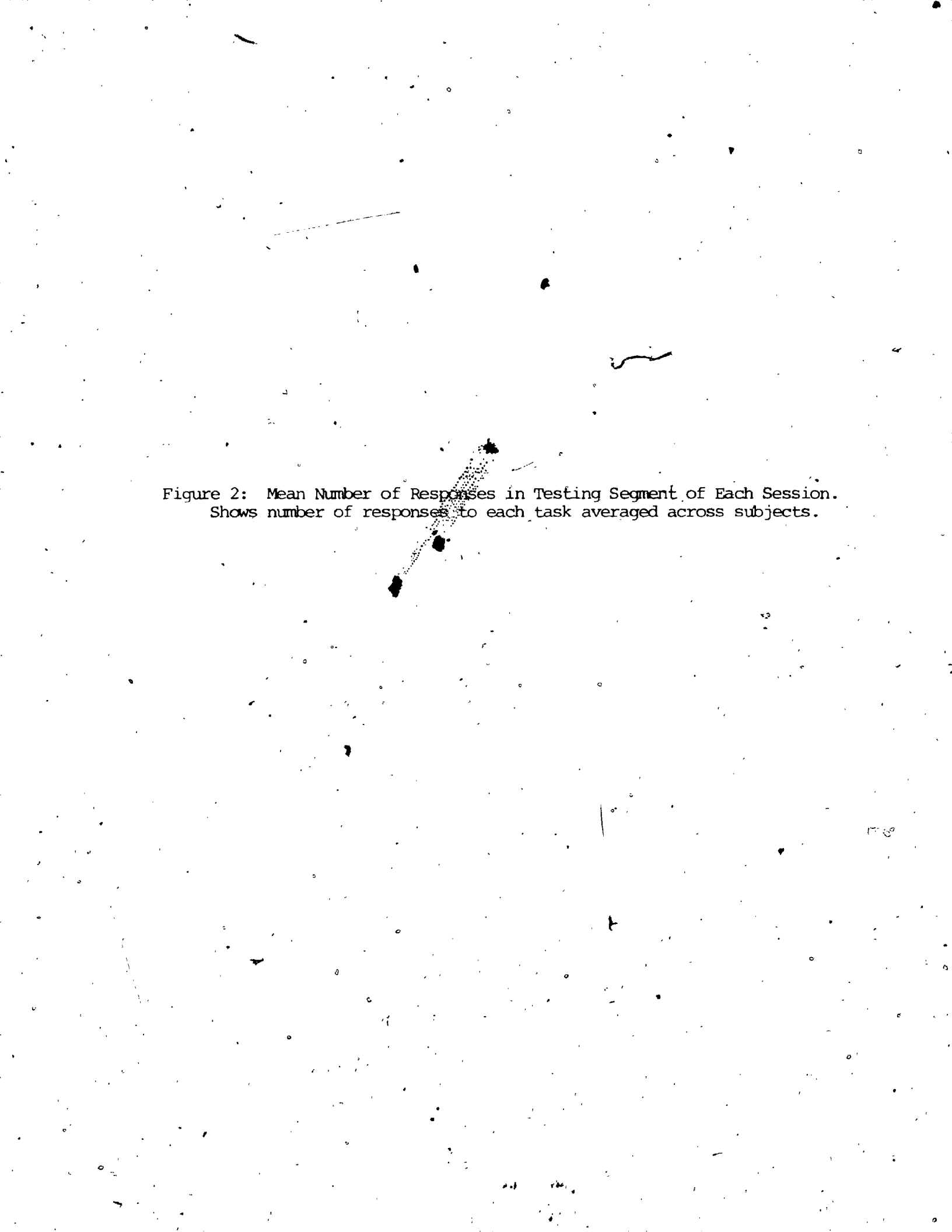
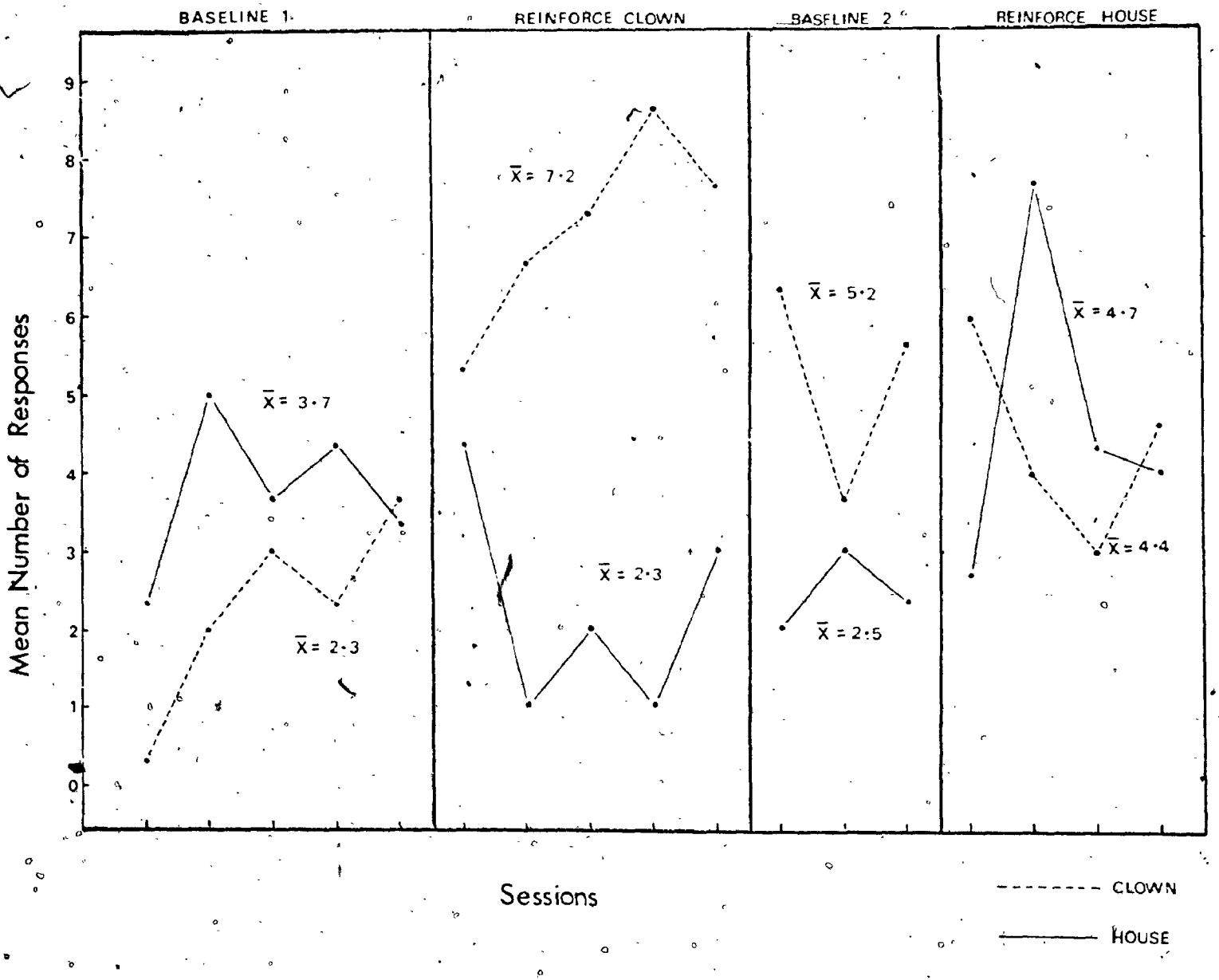


Figure 2: Mean Number of Responses in Testing Segment of Each Session. Shows number of responses to each task averaged across subjects.


BASELINE 1
REINFORCE CLOWN
BASELINE 2
REINFORCE HOUSE
Mean Number of Responses
0
1
2
3
4
5
6
7
8
9
X̄ = 3·7
X̄ = 2·3
X̄ = 7·2
X̄ = 2·3
X̄ = 5·2
X̄ = 2·5
X̄ = 4·7
X̄ = 4·4
Sessions
CLOWN
HOUSE

The above data give no indication that responding to either task in Testing segments decreased when the task was reinforced in Training segments, regardless of whether that responding is compared to the preceding Baseline or to the unreinforced task responding. In fact all comparisons indicate some increase in Test segment responses to the reinforced tasks during Reinforcement.

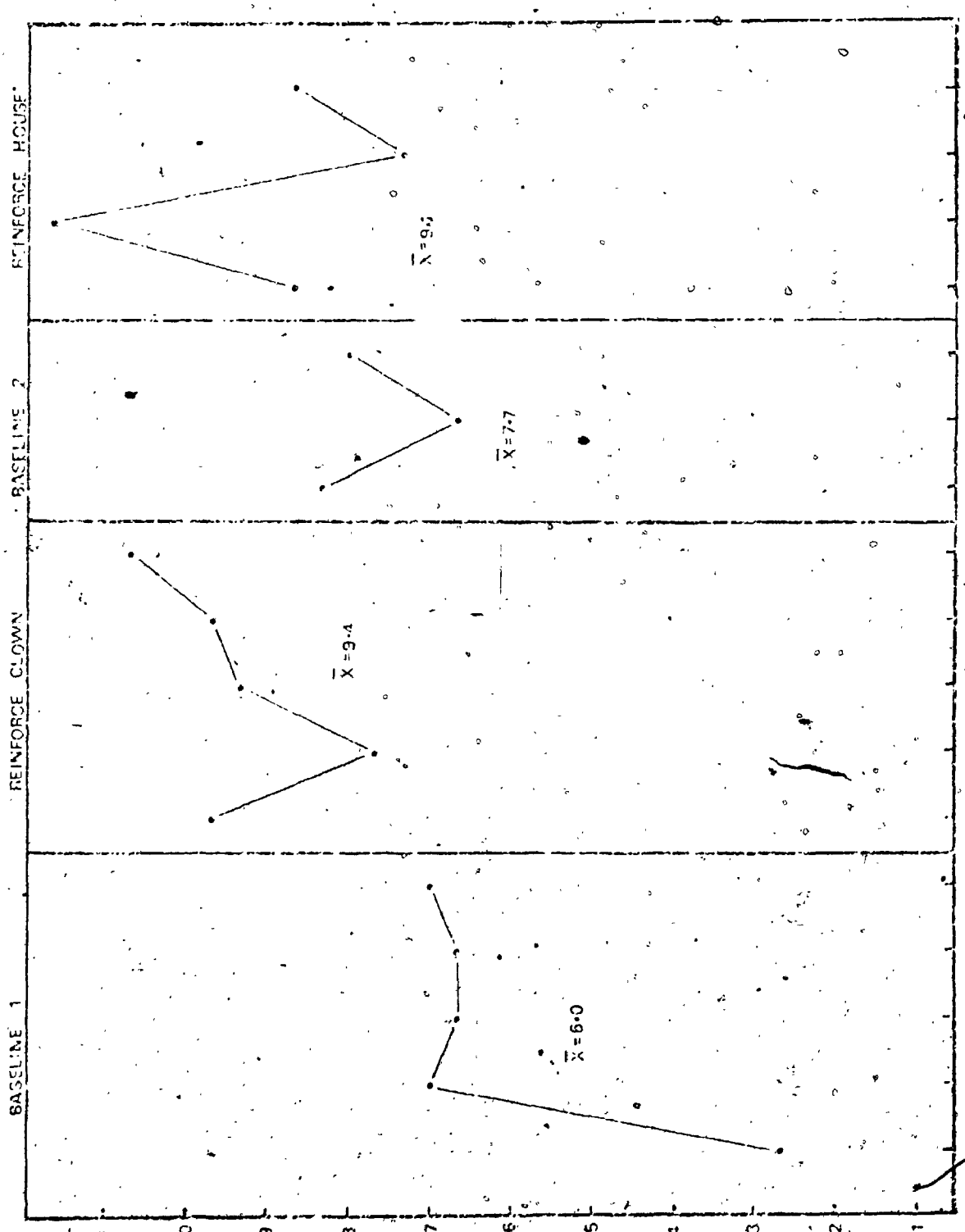
It is possible that this failure to demonstrate the Overjustification Effect could be due to an overall decrease in responding to both tasks when one was reinforced, indicating a general decrease in interest in both tasks. To determine if this was the case, responses to the machines in each session were summed across machines and across subjects. Since responses were counted for a fixed five minute interval each day, it was assumed that the overall total number of responses combined for both machines would reflect changes in overall interest level from session to session. More specifically such figures should indicate whether changes in responding under reinforcement conditions were due to shifts in preference, or to some change in overall interest in both tasks. Figure 3 shows the mean total responses per day, calculated by summing the responses made by all subjects to both machines, and dividing by the number of subjects. The mean rate of response by condition was 6.0 in

---

Insert figure 3 about here

Baseline 1 and 7.7 in Baseline 2. It was higher during both

Figure 3: Mean Total Responses in Testing Segment of Each Session. Shows total responses to both Clown and House Tasks per session averaged over subjects.


BASELINE 1
REINFORCE CLOWN
BASELINE 2
REINFORCE HOUSE
$\bar{X}$=6·0
$\bar{X}$=9·4
$\bar{X}$=7·7
$\bar{X}$=9·3
Mean Total Responses
Sessions
1
2
3
4
5
6
7
8
9
10
11

reinforcement conditions (9.5 during Reinforce Clown and 9.1 during Reinforce House).

Training Segments: Reinforcement was delivered contingent upon any response to the designated machine, and was independent of speed or accuracy of the response. This was done to ensure that there would be little difference in responding during Training as a result of differential reinforcement.

Subjects were all presented with an equal number of Training trials on the two machines in each session. Table 2 shows the mean number of errors made by each subject in each condition

---

Insert table 2 about here

---

by task, and the mean difference in response time for tasks in each condition. The number of errors made was very low, and accuracy always exceeded 85%. No systematic differences in error rate between tasks occurred with the onset of reinforcement. The mean difference in response time was calculated as follows: the average time taken from initial contact with the card by a subject until he pressed a button on the machine was calculated for each task in each session. The value thus obtained for the Clown task was subtracted from the value for the House task. The mean of these differences was then calculated for each condition for each subject. It was predicted that if reinforcement were to have any effect on the speed of response, this mean difference value should be different from Reinforce House to Reinforce Clown as reinforcement was applied to different tasks. This was tested using a

Table 2: Number of Errors and Difference in Time of Response. Shows number of errors for each subject, averaged for each task in each condition. Also shows the difference in average time per response between the two tasks for each subject in each session, averaged within conditions.

DAY	SUBJECT 1			SUBJECT 2			SUBJECT 3		
	ERRORS		TIME	ERRORS		TIME	ERRORS		TIME
	House	Clown		House	Clown		House	Clown	
1	0	2	-1.1	1	3	6.5	0	0	NO DATA
2	0	1	5.8	1	1	6.3	0	1	-2.2
3	0	1	0.1	0	0	2.0	0	1	0.6
4	0	1	-1.6	0	2	5.2	0	0	-5.8
5	0	1	-2.0	0	2	-0.9	0	0	-13.7
$MEAN_1$	0	1.2	0.2	0.4	1.6	2.4	0	0.4	-5.3
6	0	0	-1.3	0	1	-4.2	0	0	-1.1
7	0	2	1.9	1	0	2.1	0	1	-2.5
8	0	2	0.0	0	0	-0.5	0	0	-1.4
9	0	0	-0.2	0	2	-3.3	0	0	7.7
10	0	0	-0.7	0	2	-0.5	0	0	6.8
$MEAN_2$	0	0.8	-0.1*	0.2	1.0	-1.3**	0	0.2	1.9***
11	1	0	-6.5	0	2	-6.7	0	0	7.9
12	0	0	-12.0	0	1	-0.7	0	0	-7.1
13	0	0	-5.7	1	0	-5.3	0	0	-6.0
$MEAN_3$	0.3	0.0	-8.1	0.3	1.0	-4.2	0	0	-1.7
14	0	2	-3.2	0	1	2.0	0	0	3.6
15	0	0	-2.3	0	0	1.0	0	0	-10.2
16	0	0	1.4	0	2	-1.3	0	0	-1.1
17	0	0	6.5	0	0	-0.9	0	1	-25.7
$MEAN_4$	0.0	0.5	0.6*	0.0	0.8	0.9**	0.0	0.3	-8.4***

* $Mean_2 - Mean_4 = -0.54$ ($t = 0.059$, $p > .10$)
** $Mean_2 - Mean_4 = 0.43$ ($t = 0.100$, $p > .10$)
*** $Mean_2 - Mean_4 = 4.78$ (t 0.517, $p > .10$).

one-tailed t-Test for independent means.[3] The difference between the two means was not significant for any of the three subjects ($p > .10$), indicating no systematic difference in response time for conditions.

Self Report Data: Subjects were asked, at the end of sessions, to indicate which machine they liked best. If the subject made no response, or appeared not to understand, this was expanded: "If you could take one of the machines home with you, which one would you take?" After the first Training session of Reinforce Clown, subjects were also asked which machine they were earning tokens for.

On 75% of the reinforced sessions, subjects reported preferring the machine which was currently being reinforced in Training sessions. Subjects named various performance characteristics (e.g. colour, bell, music, etc.) as the reasons for their preferences most of the time. However S_1 attributed his verbalized preference to the fact that he got tokens for it a total of three times out of nine reinforced sessions, and S_2 did the same a total of six of the nine reinforced sessions. S_3 never verbalized that the machine preference was due to tokens.

Subjects made no errors in identifying which machine they were getting tokens for, and all subjects agreed each day to perform the reinforced task in order to get tokens.

Reinforcer Effectiveness: This was determined during the Baseline 2 condition, and during the final session of the study. The index of effectiveness used was the percentage

of responses made to the reinforced hole (S^+) after the first reinforced response was made. The mean of this value over subjects equals 77% (range=66% to 100%). The mean for S_1 was 90%, and for S_2 and S_3 the mean was 70%.

DISCUSSION

This experiment demonstrates that extrinsic positive reinforcement of responding in one situation can influence children's choices of activities in a different, free-play situation; but here, contrary to the predictions of the Overjustification Hypothesis, reinforcement led to increased choice of the previously reinforced activity, indicating increased rather than decreased interest in the task. The effect was shown to persist over several sessions. Now, since the experiment was designed to meet the conditions deemed necessary for observation of the Overjustification Effect (Greene and Lepper; 1975; Lepper and Greene, 1976) the failure to demonstrate that effect here raises serious questions about the applicability of the Overjustification Hypothesis to ongoing reinforcement systems.

In considering this study it is necessary to examine the actual findings, and the applicability of those findings to the Hypothesis.

The Effects of Reinforcement on Responding

There are three areas where the effect of reinforcement was examined in this study:

i) Direct effects on Clown and House tasks in Training Segments: Practice effects arising from reinforcement procedures have been confounding factors in many studies in

this area. In the present study reinforcement procedures had little or no effect on responding in Training segments.[4] The data show that the administration of reinforcement did not significantly affect the rate of response for any subject in Training segments, and there is no evidence of a systematic change in accuracy on Training Trials with either the onset or offset of reinforcement. In addition, subjects made an equal number of responses to each machine in each Training segment.

With the effects of practice thus minimized, the effects of reinforcement on responding in the Testing situation (Interest) can be examined.

ii) Direct effects on marble task in Training Segments: Since the reinforcers delivered in Training segments had no effect on responding in those segments, it might be argued that they were not reinforcers at all, and that subjects did not perceive themselves as being rewarded. However, the data on responding to the marble task indicate that when the same reinforcers were delivered in the same situation for a different task, a reinforcement effect was demonstrated. All subjects came to choose to drop marbles into the side for which tokens were given, on an above chance basis.

The reinforcers employed were thus demonstrated to be effective reinforcers in the same situation where Training took place. Direct evidence of effectiveness with the specific tasks under study was not obtained, since such a demonstration would have differentially affected practice on the two tasks,

and thus confounded the experiment. (see also footnote 4)

iii) Indirect effects on Clown and House tasks in Testing segments (interest): The present design incorporates a multiple baseline design with aspects of an ABAB design. Subjects' responding to different unreinforced tasks during Reinforcement Testing segments serves as one comparison for responding to tasks reinforced during Training segments, while their responding to the same tasks in different conditions serves as another comparison.

The onset of the Reinforcement condition for one task in Training segments led to an increase of choice of that task in the corresponding Testing segments, while leading to a slight decrease in choice of the unreinforced task. Since increased choice was specific to the reinforced task, it can be concluded that reinforcement of a specific response caused increases in that response in Testing segments.

The offset of reinforcement for the Clown task in Baseline 2 led to a decrease in choice of that task in Testing segments relative to House task responses in the same condition, strengthening the conclusion that reinforcement procedures in Training segments caused the changes in Testing segment responding, since only the previously reinforced response was affected.

The relative response preference data shows that in both Reinforce conditions, the reinforced task became relatively more preferred than it was in the preceding Baseline. This constitutes a demonstration and replication of increased interest due to reinforcement.

These results clearly do not support the Overjustification Hypothesis.

The Validity of the Procedure

Lepper and Greene (1976) and Greene and Lepper (1975)

have outlined conditions which they deem necessary but not sufficient for the Overjustification Effect to be observed, and it might be argued that the present study failed to demonstrate the effect because it did not meet these conditions. Thus the conditions will be examined one by one, and the extent to which the present study has met each will be discussed:

i) Continuing instrumentality: If subjects were to erroneously perceive that the tasks which they are performing might have continuing instrumentality (might continue to lead to extrinsic reward) even after the reinforcement were withdrawn, this could hide the Overjustification Effect. In the present case however, the Training and Testing segments were designed to be maximally discriminable. In Training segments, where reinforcers were delivered, an experimenter was present, a selection of candies and trinkets was present, tokens were present, and subjects were instructed that they would be given tokens for a certain response; whereas in the Testing segments none of these was the case. In addition, subjects were never given rewards immediately following a Testing segment, and the contingencies of reinforcement were carefully explained each day. Moreover, Testing segments always took place at least 23 hours after Training segments minimizing the possibility of carry-over of recent experience.

The fact that sessions were carried out in the same

room may be argued to have enhanced the possibility that subjects would perceive their responses to have continuing instrumentality. This is a remote possibility since no reinforcement was delivered in any of seventeen identical Testing segments. In addition, Ross, Karniol and Rothstein (1976) demonstrated that the Overjustification Effect can occur even if testing is carried out in the same room where reinforcement was delivered and with the person who delivered the rewards present during Testing. Thus it is unlikely that the present failure to observe the effect is due to the fact that Training and Testing took place in the same physical surroundings.

It is, however, impossible in this study to prove that subjects did not perceive the contingencies as continuing, and there is ample evidence that contingencies no longer in effect can affect behaviour over many sessions (e.g. Bucher and Davidson, 1974). While we have no direct evidence that subjects did not perceive some sort of continuing instrumentality in the Testing segments, the fact that responding was maintained over several Testing segments, when each day this perception was disconfirmed by the fact that no reinforcers were received in Testing segments, suggests that no such perception existed.

ii) Information about competence: The reinforcers used in this program did not provide subjects with any information about their competence at the task. Reinforcers

were dispensed contingent upon the response of pressing one or the other of the buttons, regardless of skill or accuracy. Moreover, the verbal information about correctness of the response was not a possible source of the failure to show the effect for two reasons: a) it was redundant with the information automatically emitted by the apparatus. b) it was not confounded with reward, but administered in all Training segments for both tasks.

iii) Skill development: For the Overjustification Effect to be observed, subjects must not increase in skill at the reinforced task due to the reinforcement, since such increased skill might lead to changes in subjects' perceived level of competence and thus increase interest. In the present study, the tasks chosen were simple for subjects with initial accuracy being between 87% and 100% during Baseline. Moreover there is little likelihood that subjects' skill would increase during reinforcement since this was delivered regardless of correctness of the Training response. While feedback about accuracy was delivered both verbally and automatically, this occurred in all sessions and did not covary with reinforcement.

As can be seen in Table 1, subjects initially made more errors on the Clown Task than on the House Task, and gradually improved through the study. Some improvement was made during conditions where reinforcement was not delivered for that task (e.g. Baseline 2), and under these conditions of non-reinforcement improved skill was not

reliably accompanied by increased choice of the Clown Task. It is thus concluded that skill development is a negligible factor in this study.

The whole question of skill development and Overjustification in token economies is an important issue which is discussed later in this paper.

iv) Salient rewards: Greene and Lepper (1975) point out that for the Overjustification Effect to be observed, the rewards used must be salient or "psychologically visible". While there is some debate as to the definition of salience in this context, the reinforcers used in the present experiment seem, by most definitions, to be salient.

One possible definition of salience might involve the effects of reward on behaviour, and here tokens were shown to be effective reinforcers in the Marble task. However, Lepper and Greene (1976) argue that reinforcer effectiveness is independent of salience, and suggest that salience should be inferred from the reward procedure (e.g. Ross, 1975).

The procedures used in this experiment were designed to promote the salience of the reinforcers. These items were prominent during Training and subjects assisted in carrying them to the room. Subjects thus touched and had visual access to them in each session.

An important aspect of salience is that subjects perceive the reward as being contingent solely upon the reinforced task. To determine this the experimenter asked subjects to identify which machine was the S^+ in each session, and subjects invariably answered correctly. Moreover the instructions at the outset of each session noted which machine led to tokens, and while two subjects reported on several occasions that their preference for the reinforced task was due to tokens, no subject ever

indicated that tokens were being delivered for the non-reinforced task.

In light of the manner of administration of rewards, together with the effects which they had on both verbal and motor responses, it is concluded that the rewards used and the contingencies in effect were salient to subjects.

Relevance of Methodology

Typical token economy research has been dismissed by Lepper and Greene (1976) as being irrelevant to the Overjustification Hypothesis because of three features lacking in that research methodology. The present study is discussed below in terms of each feature:

i) Initial Interest: For the Overjustification Effect to be observed, it is necessary to demonstrate that the task is of some initial interest. This brings up definitional problems. If one is to accept the concept of scientific determinism, one has to assume that initial interest is determined, at some point, by elements outside the organism. Greene and Lepper (1975) and Lepper and Greene (1976) have recognized this fact in admitting that interest is situationally specific. Thus a response which has a high probability of occurrence in a nursery school may have a low probability of occurrence in a candy shop. Lepper and Greene indicate that under a specific set of conditions, if subjects emit a response at a high rate to the exclusion of competing responses, interest may be inferred.

Interest is thus determined by relative rather than absolute reponse rate. In the present study all subjects made some responses to both tasks during Baseline 1. The fact that both responses occurred may indicate something about initial interest. However the relative rate of response to both tasks is more important in discussing the effects of reinforcement. Decreases in interest in either task could have been accompanied by either an increase, a decrease or no change in the other task. Use of relative rate of response takes changes in responding to the other task into account.

If absolute rate were employed it could be argued that experimenter demand was a factor in the initial interest shown in the tasks since the experimenter questioned the subject about his/her responses and verbally encouraged subjects to play with the machines during Testing. However experimenter demand is not really an issue here since, while subjects were constrained to some extent to make responses in Testing segments, there were no constraints placed on choice of tasks, and thus relative interest is not affected.

Interest was determined in a similar fashion by Greene (1974) where students were initially asked to choose one of a small set of responses, and then changes in interest due to reinforcement of some of the responses were gauged by changes in relative responding. In that study, as in the present one, subjects were expected to respond, but choices between responses were unconstrained.

Since subjects initially emitted responses in the situation in the absence of either reinforcement or punishment, and since the measure of interest employed was similar to that of Greene (1974) it is concluded that subjects showed initial interest in the activities suitable to allow for a test of the Overjustification hypothesis.

ii) Appropriate controls: Lepper and Greene (1976) assert that most token reinforcement studies lack appropriate controls necessary to make inferences about post reinforcement response levels. The present experiment was designed so that responding in the Testing segments could be assessed with each subject's previous behaviour serving as one control. In addition there was a second task for which no reinforcement was dispensed, which controlled for individual variations from session to session. Despite this, there was no indication that the Overjustification Effect occurred.

The possibility existed that total responding to both tasks could have been lower after reinforcement indicating an overall loss of interest in the pair of activities. However, figure 3 shows that this is not the case, and that overall responding during Testing segments during reinforcement conditions was higher than during non-reinforcement phases, supporting the conclusion that responding was increased by the reward procedures.

iii) Freedom from potential controlling agents: In typical reversal designs used to investigate token economies, there is not a significant change in the setting between the reinforcement phase and the return to Baseline conditions. Often the same people are present (teachers or ward staff) and these individuals may, through their actions or presence, influence subjects' responding.

In the present experiment, Testing was carried out in the same apparatus. However, there were no potential controlling agents there. The tokens, prizes, and experimenter were all absent from the room and from sight. Moreover, the experimenter led the child to believe that he was busy with another job (making orange juice) and not concerned with the subject's performance.

Summary

The present experiment attempted to meet all the conditions for a fair test of the Overjustification Hypothesis. In meeting these conditions, the study was somewhat different from most reinforcement systems, since it reinforced a high frequency behaviour, and did not try to increase speed or accuracy. The fact that the Overjustification Effect did not occur here brings into serious question the assertion that Overjustification may pose a problem for ongoing reinforcement systems.

The present study constitutes a demonstration that interest in tasks, as inferred from unconstrained choice of tasks, can be increased by the use of token reinforcement backed by tangible rewards. The tasks, while simple for the subjects, were academically important tasks involving the identification of numbers, letters and pictures. The reinforcement system used was typical of token systems recommended by behaviour analysts, in which the reinforced behaviour was specified, and consistently consequated following

its occurrence. The administration of tokens was frequent, and systematic.

It must be pointed out that the experiment took place in a special setting where children of normal intelligence were dealt with on a one-to-one basis, and thus is open to the criticism of being an analogue. The study which follows is a similar investigation which was conducted in a naturalistic setting.

EXPERIMENT 2

Investigations into the effect of reward upon intrinsic interest have primarily been analogue studies, in some instances assessing the effects of reward using analogue academic tasks (e.g. Reiss et al., 1975; Feingold et al., 1975) and in other instances using both academic tasks and reward systems which are analogues of actual applications (e.g. Lepper et al., 1973; Salancik, 1975). The relevance of such analogue studies to actual applied questions is in doubt (Reiss et al., 1976).

The purpose of the present study is to investigate the effects of reinforcement in a token system upon interest in an academic task within an applied setting. Despite the criticisms of the Overjustification Hypothesis, many authors have generalized the findings of analogue studies which support it, to ongoing token systems in the classroom. In light of the lack of specified conditions in the hypothesis, these authors have assumed that the critical conditions which determined the results in the analogue studies also exist in the classroom token system. While it is apparently appealing to make such generalizations, their empirical basis is limited. There exists disagreement within the analogue studies as to whether the effect occurs or not, and there is no strong empirical support from actual applied

situations to justify such generalization.[5]

The present experiment takes place in a classroom setting, and two hypotheses are put forth:

Hypothesis I: That the administration of an ongoing token reinforcement system, within an actual classroom situation, with multiple reinforcers being delivered contingent upon academic responses will not measureably decrease subjects' response rate following the withdrawal of the token system. This is essentially a null hypothesis, but, in light of the lack of support for the Overjustification Hypothesis, failure to reject it would have serious implications regarding the generality of the conditions under which the Overjustification Hypothesis applied.

Hypothesis II: That administration of reinforcement will lead to a measurable increase in the response rate of subjects in those academic responses, following the withdrawal of reinforcement, compared to Baseline.

In order to maximize the probability of the occurrence of the Overjustification Effect, the conditions under which Greene and Lepper (1975) and Lepper and Greene (1976) have indicated the effect may occur have been met as nearly as possible by this study.

Support for Hypothesis I or II would not disprove the Overjustification Hypothesis, but would simply indicate that the conditions necessary for the effect to be observed do not occur in this study. The question being addressed here is whether or not the Overjustification Effect occurs

in properly run token systems in applied settings as is predicted by the Overjustification Hypothesis (Lepper and Greene, 1976; Greene and Lepper, 1974; Levine and Fasnacht, 1974; Notz, 1975). The design is such that the Effect should not be hidden by aspects of the procedures employed, a problem cited by Lepper and Greene (1976) as rendering most of the token economy literature irrelevant to the Overjustification Hypothesis.

METHOD

Subjects

Subjects were 59 full time students enrolled in an undergraduate course in the Psychology of Exceptional Children (Psychology 342) at the University of Western Ontario. Of these subjects, three dropped the course after the first session, while the rest remained throughout the experiment. Students were divided into three groups by the Registrar's office of the univseristy, on the basis of the hour at which the group took the class. The three groups met in three consecutive tutorial classes on the same day every week. Group 1 consisted of 19 students; group 2 of 16 students; and group 3, of 21 students.

Setting

The study took place in the regular classroom. Briefly the course operated as follows: Students were assigned approximately forty pages of readings each week, whose emphasis was upon techniques for intervention with exceptional children. Students were supplied weekly with "cue sheets" which highlighted the readings page by page, giving key word references to important passages. In tutorial sessions, students were called upon to discuss the readings, and in preparation for this, they were requested to hand in, at the beginning of each tutorial session, a list of the key word cues from the cue sheet which they were prepared to discuss, together with a brief written discussion of each cue listed (response). Students

achievement in the course was not determined by the responses handed in, but rather by grades on weekly tests covering a subset of items on the cue sheet, together with grades on formal papers and projects. In addition a small portion of the grade was determined by attendance at tutorials.

The weekly test took place the day after each tutorial session, when all three groups met at the same time and location for a class. The test involved a small number of short answer questions, which were completed and marked by students in one session. Marks were recorded by the instructor during the following week, and the test paper returned to the student at the next tutorial session, a week later.

Psychology 342 had been conducted according to this format for two years prior to the present experiment, and this study investigates a manipulation made within the ongoing format.

Procedure

Reinforced response: In order that conditions outlined by Greene and Lepper (1975) might be met, the response to be reinforced was selected on the following basis: a) that it was of an initial free operant level greater than zero; b) that it be such that practice and feedback would not lead to increased skill in other areas, and thus to secondary gains for subjects. The response selected was the students' written discussion of key word cues handed in at the first of each tutorial session. The response was under control of experimenter demands to some extent, in that subjects were

asked to hand in some cues. However the number of cues to be handed in was never specified. To avoid skill training reinforcement was not delivered for the content of the responses to the key word cues, but for the number of key word cues responded to.

Design: The experiment was designed to run over one semester, with 11 weeks of data collection. This time period was divided into three experimental conditions.

i) Baseline 1: For the first three sessions the initial rate at which responses to cues were handed in by students was determined. All groups were treated alike during Baseline. In the course description students were advised to hand in from one to 10 responses to key word cues each week. The week after a set of responses was handed in, the student was informed of the number of responses he had made by a written message at the bottom of the test for that week, which was handed back at the subsequent tutorial. If a student had handed in no responses the feedback was: "Number of cues handed in - 0". During Baseline students received one point counting toward the final grade, contingent upon attendance at the tutorial class.

ii) Reinforce Cues: In this condition the three groups were treated differently. At the end of the third session students were given a cue sheet which informed them of the specific conditions which applied to their particular group. The assignment of groups to conditions was carried out by a member of the secretarial staff who was not informed of the

design of the experiment or the composition of the groups, and who distributed one set of instructions to each group. The instructor was not informed which of the treatments the specific groups had received.

All groups received a set of common instructions followed by instructions specific to their group. The common instructions read:

> I plan to do an experiment next term on the value of the cue sheets and class procedures I use in this class. As a preliminary, I want to get some information this term about the tutorials. Each tutorial will be treated somewhat differently. This procedure introduces the possibility of contamination within the tutorial groups which I want to avoid. The procedure for this tutorial is described below. I can't talk about it in class, since I have arranged it so I won't know which tutorial gets which procedure, unless you tell me (so please don't). This way I won't be biased, and the class as a whole won't be affected by class discussion of the different procedures.

Subjects were informed that they were in the experiment only by their consent, and were told that if they did not wish to participate, that they should contact the teaching assistant in the course who would make arrangements for them. They were told that this was a preliminary study and

would last only a few weeks. No student withdrew from the study.

In addition to the general instructions, each group received instructions applicable for that group.

a) Group 1 (Control). The control group received:

> For your group there will be no change from the previous procedure. The rest of the space I am typing in just for filling, so all the special messages will be the same length as messages handed out in the other tutorials.

b) Group 2 (Reinforce Increase). This group was given instructions intended to increase the number of responses to cues handed over base rate:

> For your group you will have to earn your tutorial point by the cues you write and hand in. This may help you prepare for the exam. You'll get one seventh point for each cue you hand in each week, up to one point per week.

c) Group 3 (Reinforce Decrease). This group was given instructions which were intended to decrease the number of responses handed in compared to Baseline.

> For your group you will have to earn your tutorial point by the cues you write and hand in. This may help you prepare for the exam. You'll get one fourth point for each cue you hand in each week, up to one point per week.

In Baseline 1, the median number of responses handed in each week across groups was between five and six. Thus, for group 2, seven responses was selected as the maximum to be reinforced since it constituted an increase over the median during Baseline. For group 3, four responses was selected as the maximum to be reinforced since this constituted a decrease over the median during Baseline.

Apart from the handouts no explanation for the experimental manipulation was made. Any questions posed to the instructor were turned away with an explanation that he did not wish to discuss the manipulation for fear of experimental contamination.

iii) Baseline II: At the end of four sessions of Reinforce Cues, students in all groups received the following message on their cue sheet:

> The experimental procedure I introduced a few weeks ago has now ended. Beginning with the cues below, we will return to the preceding system of awarding points at tutorials. Points will be given for attendance. As before, I would like you to continue handing in cues each week, from one to 10 as before.

Students continued throughout Baseline II to get feedback each week as to the number of cues they handed in the week previous, but apart from that, the manipulation was never again mentioned. Baseline II continued for four sessions.

Reinforcement: The reinforcement in this study was a point or part of a point delivered in each session, counting toward the final grade of the course. In the two Baseline conditions the point was delivered contingent upon attendance, while in the reinforcement condition (Reinforce Cues) the point was divided up for the two experimental groups, and delivered contingent upon responses handed in, up to a specified maximum. This procedure was designed to equate for reinforcer density across groups and conditions.

After the experiment was concluded, points were re-calculated for tutorial sessions. For ethical reasons points for all groups were re-calculated on the basis of attendance at tutorials only, disregarding the number of responses made. Thus while each group was treated differently for the experiment, this difference was removed after the experiment was concluded, and the responses had been recorded. Subjects were debriefed following the last session, informed about the general design of the experiment, and the changes in the points for tutorial sessions.

RESULTS

Response Definition

The number of responses to cues handed in by subjects constituted the primary data for this study. A response to a cue was counted if the student listed the original key word cue from the handout, and added one or more words of expansion. For the most part students numbered or otherwise differentiated among the responses they handed in, and this number was accepted if all the cue responses met the loose criteria above.

Reliability

Each week, each student's responses to the cues were counted by a rater, and this number was reported back to the student the following week. For one session in each of the three experimental conditions, another independent rater was instructed to count the number of cues handed in by each student, as a check of inter-rater reliability. This rater, who was unaware of the experimental design, first studied the handout for the session in question, then read the sheets handed in by each student, and recorded the number of responses for each student.

Reliability was calculated as the total number of papers where the two raters were in exact agreement as to the number of responses, divided by the total number of papers handed in for the session. For the purposes of

determining reliability, only students handing in one or more cues were included, so as to prevent zero scores from inflating the reliability estimate.

Inter-rater reliability on the number of cues handed in was high (session 3: 0.90, session 5: 0.96, session 10: 0.92) indicating that the feedback received by students was accurate.

Number of Responses Handed In

The task reinforced was the handing in of responses to cues. Two of the groups received reinforcement for handing in at least a fixed number of responses, different for each group. In order to assess the impact of this manipulation the difference between the mean number of cues handed in before and after treatment was compared across groups.

The mean number of responses handed in was calculated for Baseline 1 and Baseline 2, in each case by dividing the total number of responses which a student made in each condition by the number of sessions in that condition. The number of sessions could be calculated in two ways: either as the number of sessions in the condition, or as the number of sessions the student attended. The former method decreases the estimate of means for subjects who fail to attend a session (in effect scoring zero response for such a session (Absentees Equal Zero)), and this decrease might reflect a decrease in interest, or be confounded with other

variables and simply add undue variance to the estimate of interest. The latter method (Ignoring Absentees) avoids this. Since arguments can be presented favouring both methods of calculation, both were used.[6]

For each subject the mean number of responses in Baseline 1 was subtracted from the mean in Baseline 2. An analysis of variance on the difference scores thus obtained revealed no significant differences among the groups, regardless of the treatment of absentees (See figure 4 and table 3).

Insert figure 4 and table 3 about here

All groups showed a decrease in responses from Baseline 1 to Baseline 2, but this decrease was least (in terms of absolute numbers) for the Increase group under both analyses.

Proportion of Sessions Where Responses Made

In an effort to reduce the within group variance the proportion of sessions where subjects made one or more responses to the cues was calculated for each subject in each group. Thus for each session subjects were scored either 1 or zero, and these values were averaged within conditions. This transformation had the effect of reducing the data for each session to binary form. As above the data could include or exclude sessions which the subject did not attend, and

Figure 4: Mean Number of Responses per Subject Per Session. Shows means session by session and averaged within each condition, using two approaches to absentees.

FIG. 4

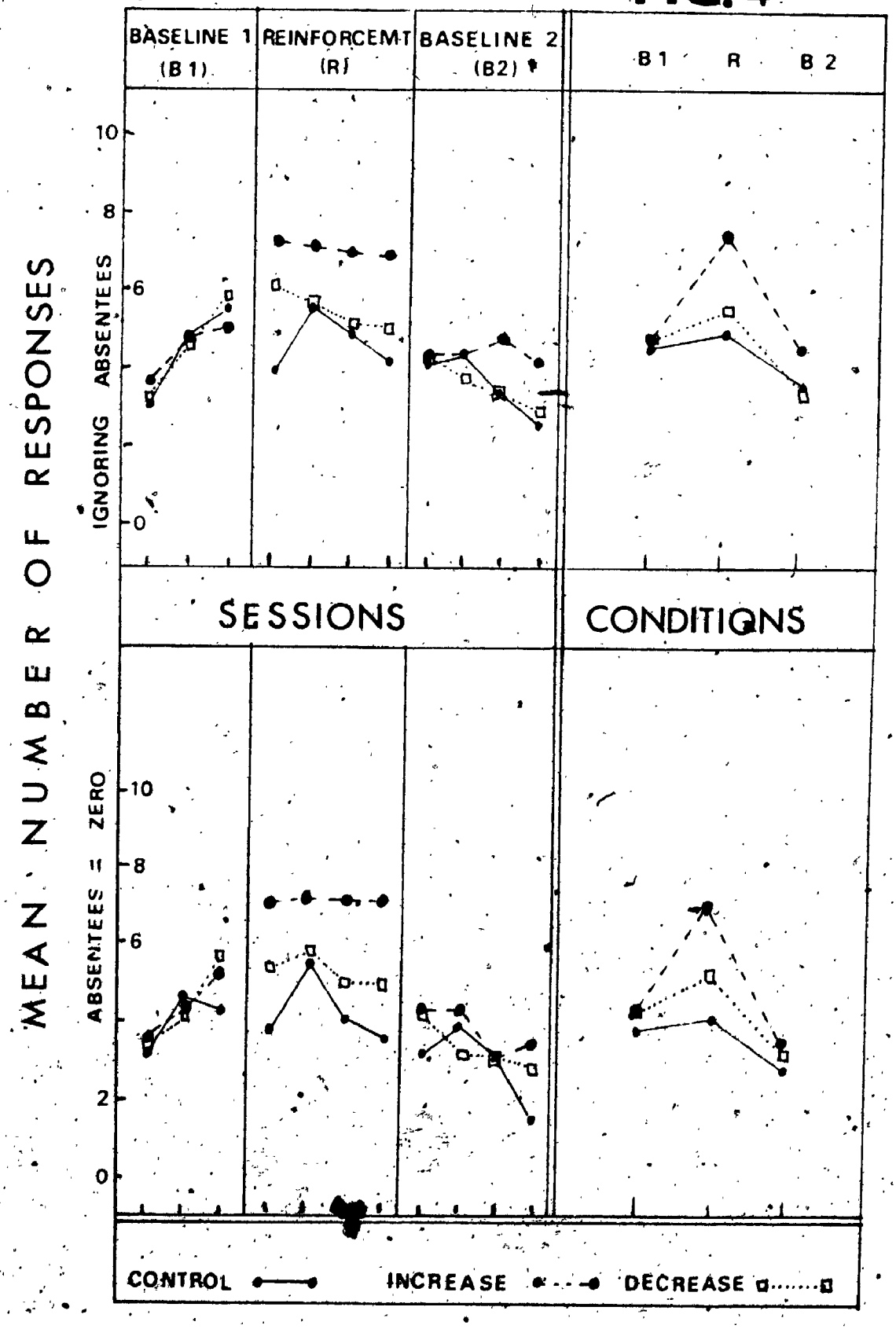

BASELINE 1 (B1)
REINFORCEMT (R)
BASELINE 2 (B2)
B1
R
B2
MEAN NUMBER OF RESPONSES
IGNORING ABSENTEES
ABSENTEES = ZERO
0
2
6
8
10
SESSIONS
CONDITIONS
CONTROL
INCREASE
DECREASE

TABLE 3: Comparison of Differences Between Mean Number of Responses During Baseline 1 and Baseline 2 Across Groups.

Calculation Using Data Ignoring Absentees:

Group	Mean Number of Responses Baseline 1	Baseline 2	Mean Difference
Control	4.19	3.67	-0.52
Increase	4.53	4.52	-0.01
Decrease	4.59	3.48	-1.11

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	27.46	2	13.73
Within	318.08	53	6.00
Total	435.54	55	

$F = 2.29$; $df = 2, 53$ $p < .05$

Calculation Using Data Counting Absentees as Zero:

Group	Mean Number of Responses Baseline 1	Baseline 2	Mean Difference
Control	3.93	2.91	-1.02
Increase	4.34	3.77	-0.57
Decrease	4.34	3.18	-1.16

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	2.98	2	1.49
Within	268.58	53	5.07
Total	271.56	55	

F 1.0

both calculations were used. In the case where absentees were included, they were scored as zero, and the proportion

---

Insert figure 5 and table 4 about here

---

was calculated using the number of sessions offered in a condition. In the case where absentees were ignored, the proportion of sessions where each subject responded was calculated using only those sessions where the subject was present. Because there were some absentees in each group, the latter number tended to be somewhat larger than the former for the same subject. The proportions are presented in table 4 and figure 5.

The transformation to binary data did reduce the within group variance, but did not reveal any significant differences among groups. In numerical terms the control group showed the greatest reduction of scores from Baseline 1 to Baseline 2, followed by the Decrease group. The Increase group showed the smallest decline, and, if absentees were ignored, showed an increase from Baseline 1 to Baseline 2. These differences between the groups were not statistically significant however.

Checks on Procedure

Since the manipulation failed to show any significant effect on subsequent responding, the procedure used was examined for possible explanations for this failure.

Figure 5: Mean Proportion of Sessions where Subjects Made One or More Responses. Shows proportion session by session and averaged within conditions using two different approaches to absentees.

FIG. 5

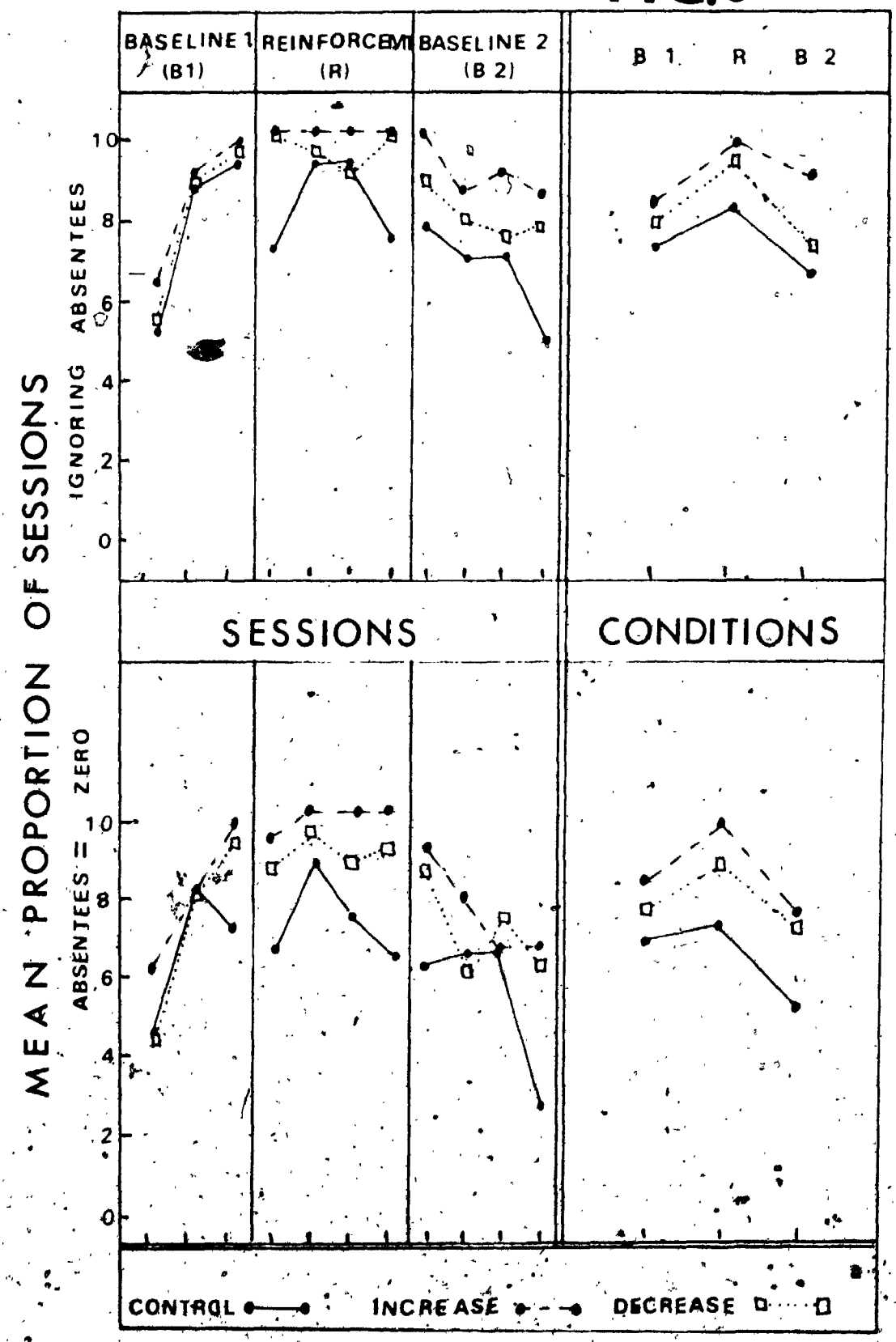

BASELINE 1 (B1)
REINFORCEMT (R)
BASELINE 2 (B 2)
B 1 R B 2
MEAN PROPORTION OF SESSIONS
IGNORING ABSENTEES
ABSENTEES = ZERO
0
2
4
6
8
10
SESSIONS
CONDITIONS
CONTROL
INCREASE
DECREASE

TABLE 4: Comparison of Differences Between Proportion of Sessions in which Responses were made in Baseline One and Baseline Two Across Groups.

Calculation Using Data Ignoring Absentees:

Group	Mean Proportion of Sessions Baseline 1	Baseline 2	Difference
Control	0.72	0.68	-0.04
Increase	0.88	0.92	0.04
Decrease	0.79	0.77	-0.02

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	0.00348	2	0.00174
Within	4.89050	53	0.09227
Total	4.89398	55	

$F < 1.0$

Calculation Using Data Counting Absentees as Zero:

Group	Mean Proportion of Sessions Baseline 1	Baseline 2	Difference
Control	0.69	0.51	-0.18
Increase	0.81	0.75	-0.06
Decrease	0.75	0.71	-0.04

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	0.20596	2	0.10298
Within	4.31965	53	0.08150
Total	4.52561	55	

$F = 1.26$; $df = 2,53$ $p > .05$

1) Secondary Gains: One possible explanation is that the Effect may have been hidden by the fact that subjects might have gained access to alternative reinforcers because of reinforcement in the second condition. It is possible that reinforcement led to higher grades in the course, and thus to higher responding to cues, even after the arbitrary external reinforcement of contingent points was removed.

This explanation cannot be supported by the data however. The grades obtained by all groups in Baseline 2 are shown in table 5. There is no difference among the

---

Insert table 5 about here

---

groups in marks obtained in Baseline 2 ($F < 1.0$).

2) Pre-manipulation Differences: Because the assignment of subjects to groups was not within the control of the experimenter, it is possible that important differences between the groups might have existed prior to the manipulation being investigated. The mean number of responses handed in by groups during Baseline 1 was compared two ways, by counting absentees as zero responses, and by ignoring absentees. In both cases the differences between groups were not significant ($F < 1.0$). (See table 6)

Insert table 6 about here

TABLE 5 : Mean Grades on Tests During Baseline 2 for All Groups.

Group	CONTROL	INCREASE	DECREASE
Mean Grade* (out of 8)	6.63	6.64	6.67

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	0.2	2	0.10
Within	52.2	53	0.98
Total	52.4	55	

F < 1.0

* Calculated by summing the grades on individual tests across sessions and across students within groups, and dividing by the number of tests marked.

TABLE 6: Initial Differences: Mean Number of Responses to Cues During Baseline 1 Across Groups.

Calculation Using Data Ignoring Absentees:

Group	Mean Number of Responses
Control	4.19
Increase	4.53
Decrease	4.59

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	1.72	2	0.86
Within	389.77	53	7.35
Total	391.49	55	

$F < 1.0$

Calculation Using Data Counting Absentees as Zero:

Group	Mean Number of Responses
Control	3.93
Increase	4.34
Decrease	4.34

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	2.07	2	1.04
Within	365.30	53	6.89
Total	367.37	55	

$F < 1.0$

The proportion of sessions in Baseline 1 where subjects made one or more responses was also compared across groups, by means of a Kruskall Wallis one-way analysis of variance, and again the differences between the groups were not significant regardless of the treatment of absentees. (see table 7)

Insert table 7 about here

The mean grade for subjects in each group in the three tests prior to the reinforcement phase were compared across groups using a one-way analysis of variance. No systematic differences between the groups were detected. (see table 8: $F = 1.25$; $df = 2, 53$; $p > .05$).

Insert table 8 about here

No significant differences existed among the groups on any of the pre-measures examined.

3) Practice Effects: The fact that the groups failed to show the differences predicted by the Overjustification Hypothesis may have been due to some differences in their responding during the Reinforcement condition, differences caused by differences in experimental conditions. One possibility is that the Increase and Decrease groups were writing more responses to cues during that period, and that this elevated response level in some way masked the effect

2 2

OF/DE

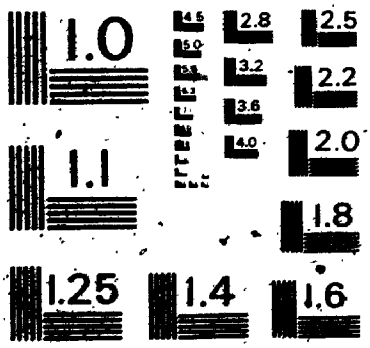


TABLE 7: Initial Differences: Proportion of Sessions in which Students Made One or More Response During Baseline 1

Calculation Using Data Ignoring Absentees:

Group	CONTROL	INCREASE	DECREASE
Proportion*	0.72	0.88	0.79

*Differences between proportions not significant ($H=2.61$, $p > .20$)

Calculation Using Data Counting Absentees as Zero:

Group	CONTROL	INCREASE	DECREASE
Proportion**	0.69	0.81	0.75

**Differences between proportions not signficant ($H=0.62$, $p > .20$)

TABLE 8: Initial Differences: Mean Grades During Baseline One Across Groups.

Group	CONTROL	INCREASE	DECREASE
Mean Grade* (out of 8)	6.26	6.38	6.68

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	1.89	2	0.945
Within	39.98	53	0.754
Total	41.87	55	

$F = 1.25$; $df = 2,53$; $p > .05$

* Calculated by summing the grades on individual tests across sessions and across students within groups. and dividing by the number of tests marked.

later. However, this possibility is not supported by the data. A t-test was used to compare changes in the mean number of reponses by the Control and Decrease during the Reinforce Cues condition. These groups were compared since the means were quite similar. For each subject the difference in mean number of responses between Baseline 1 and Reinforce Cues was calculated, and these scores were analysed. Regardless of the treatment of absentees the difference between the groups was not significant ($t = 0.51$; $p > .20$; $t = 1.42$, $p > .10$), as in shown in table 9.

Insert table 9 about here

Differences in Proportion of sessions where responses were made, from Baseline 1 to Reinforce were examined for the same two groups. These data are presented in table 10. Again a t-test revealed no significant difference between

---

Insert table 10 about here

---

the groups ($t = 0.68$, $p > .20$; $t = 1.35$, $p > .10$), regardless of the treatment of absentees.

Both groups showed increases both in responses per session and in sessions where responses were made, from Baseline 1 to Reinforce, regardless of the treatment of absentees, but no difference between the two groups was found on either measure.

TABLE 9: Practice Effects: Comparison of Differences between Mean Number of Responses in Baseline One and Reinforce Conditions.

Calculation Using Data Ignoring Absentees:

Group	Mean Number of Responses Baseline 1	Reinforce	Mean Difference
CONTROL	4.19	4.81	0.62*
DECREASE	4.59	5.50	0.91*

*Difference between differences not significant ($t = 0.51$; $p > .20$)

Calculation Using Data Counting Absentees as Zero:

Group	Mean Number of Responses Baseline 1	Reinforce	Mean Difference
CONROL	3.93	4.20	0.27**
DECREASE	4.34	5.16	0.82**

** Difference between differences not significant ($t = 1.42$; $p > .10$)

TABLE 10: Practice Effects: Comparison of Differences between proportion of sessions where responses were made in Baseline 1 and Reinforce Conditions.

Calculation Using Data Ignoring Absentees:

Group	Mean Proportion of Sessions		Difference
	Baseline 1	Reinforce	
CONTROL	0.72	0.84	0.12*
DECREASE	0.79	0.96	0.17*

* Difference between the Differences not significant ($t = 0.68$; $p > .20$)

Calculation Using Data Counting Absentees as Zero:

Group	Mean Proportion of Sessions		Difference
	Baseline 1	Reinforce	
CONTROL	0.69	0.72	0.03**
DECREASE	0.75	0.89	0.14**

** Difference between the differences not significant ($t = 1.35$; $p > .10$)

4) Insensitive Measures: One plausible reason for the failure of the present experiment to show the Overjustification Effect is that the measures may be insensitive to the change which was hypothesized to occur. Two measures were employed, and each will be examined in light of the question.

The mean number of responses was the first measure used, and this closely parallels measures used in the literature, where rate of or choice of the reinforced activity following withdrawal of reinforcement is the typical index of interest. However, it failed to detect any difference due to reinforcement from Baseline 1 to Baseline 2.

There was, however, a significant difference among the groups on this measure for difference scores from Baseline 1 to Reinforce. An analysis of variance was conducted on these scores (see table 11) which showed a significant effect

---

Insert table 11 about here

---

of reinforcement, regardless of treatment of absentees ($F = 7.68$, $p < .01$; $F = 8.81$, $p < .01$). Post hoc analysis reported above indicates that the Control and Decrease groups did not differ here. However, it is clear that the measure was sensitive to the effects of reinforcement.

Derived from the mean number of responses per session, the proportion of sessions where subjects handed in any responses was used in an effort to reduce variance. However,

TABLE 11: Sensitivity of Measures: Comparison of Differences between Mean Number of Responses from Baseline 1 to Reinforce Condition across groups.

Calculation Using Data Ignoring Absentees:

Group	Mean Number of Responses Baseline 1	Reinforce	Mean Difference
Control	4.19	4.81	0.62
Increase	4.53	7.26	2.73
Decrease	4.59	5.50	0.91

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Betweeen	44.68	2	22.34
Within	154.09	53	2.91
Total	198.76	55	

$F = 7.68$; $df = 2,53$; $p < .01$

Calculation Using Data Counting Absentees as Zero:

Group	Mean Number of Responses Baseline 1	Reinforce	Mean Difference
Control	3.93	4.20	0.27
Increase	4.34	7.16	2.82
Decrease	4.34	5.15	0.81

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	61.64	2	30.82
Within	185.58	53	3.50
Total	247.22	55	

$F = 8.81$; $df = 2,53$; $p < .01$

no effect due to reinforcement from Baseline 1 to Baseline 2 was demonstrated.

Moreover, the measure did not detect any differences among the groups in the effect of Reinforcement (comparing Baseline 1 and Reinforce differences). The summary of this analysis is presented in table 12.

Insert table 12 about here

This measure revealed no significant main effect of reinforcement on difference scores between Baseline 1 and Reinforce, regardless of the treatment of absentees ($F < 1.0$; $F = 1.16$, $p < .05$).

The mean number of responses was shown to be sensitive to the immediate effects of reinforcement, while the derived proportion scores were not.

TABLE 12: Sensitivity of Measures: Comparison of Differences between Proportion of Sessions Where Responses were Made in Baseline 1 and Reinforce Conditions, across groups.

Calculation Using Data Ignoring Absentees:

Group	Mean Proportion of Sessions Baseline 1	Reinforce	Difference
Control	0.72	0.84	0.12
Increase	0.88	1.00	0.14
Decrease	0.79	0.96	0.17

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	0.0174	2	0.0087
Within	1.8670	53	0.0352
Total	2.0112	55	

$F < 1.0$

Calculation Using Data Counting Absentees as Zero:

Group	Mean Proportion of Sessions Baseline 1	Reinforce	Difference
Control	0.69	0.72	0.03
Increase	0.81	0.98	0.17
Decrease	0.75	0.89	0.14

SUMMARY OF ANALYSIS OF VARIANCE

Source	Sum of Squares	df	Mean Square
Between	0.1804	2	0.0902
Within	2.8802	53	0.0543
Total	3.0606	55	

$F = 1.16$; $df = 2,53$; $p > .05$

DISCUSSION

This experiment was designed as a test of the Overjustification Hypothesis in an ongoing reinforcement system. The prediction derived from that hypothesis is that the Increase and Decrease groups should both show a decrease in responding from Baseline 1 to Baseline 2, relative to the control group.

The analysis of the mean number of responses per subject per session and the analysis of the proportion of sessions where subjects handed in responses both failed to support this hypothesis, regardless of the treatment of absentees. Both the Increase and Decrease groups showed a decrease, but this was paralled by a decrease within the Control group, and there was no significant difference among the groups. It is perhaps noteworthy that the Increase group always showed the smallest decrease regardless of the analysis employed.

The use of the proportion data (proportion of sessions where subjects made any responses) was an attempt to reduce variability and leave a clearer result. However the transformation to binary data made the measure insensitive to the extent that it did not detect the effect of reinforcement (see table 12). Thus the failure to show the Overjustification Effect on this particular measure is

likely due to the fact that the measure was insensitive.

However, the other measure (mean number of responses per session) has been shown to be sensitive to the immediate effect of reinforcement upon responding. Still before the failure to observe the Overjustification Effect on this measure can be discussed, it is necessary to examine some possible explanations for the failure.

i) Experimental Artifacts: The checks on procedure carried out in the result section discount several possible artifacts. Subjects in the Increase and Decrease groups did not have access to secondary gains in the form of increased grades. Moreover, practice during reinforcement is not the operative variable here, since the Control and Decrease groups were equal on this variable, and yet the predicted Overjustification Effect was not demonstrated even for the Decrease group. It was also shown that the three groups did not differ significantly on any measure at the outset of the experiment.

Thus insensitive measures, secondary gains, pre-manipulation differences, and practice effects have been discounted as possible explanations for the failure to observe the Overjustification Effect on the mean number of responses measure.

ii) Necessary Conditions: Greene and Lepper (1975) have outlined conditions which they consider to be necessary, although not sufficient for the Overjustification Effect to be observed:

a) Continuing Instrumentality. Greene and Lepper point out that subjects must view the termination of reward as just that, and that they must not perceive any continuing instrumentality in the previously reinforced activity. Figure 4 illustrates that the manipulation was effective in this regard, since in session 8, immediately following the signal that reinforcement for responses to cues had been withdrawn, the increase which had been shown by the Increase group during reinforcement disappeared.

Instructions regarding the end of reinforcement were clearly spelled out on a cue sheet. The fact that they influenced one group's responding so dramatically tends to discount the notion that a perception of continuing instrumentality was a variable masking the Overjustification Effect.

b) Skill or Competence Training. Greene and Lepper argue that if subjects come to view themselves as more competent at the reinforced task as a result of reinforcement, they will be more likely to engage in it after the reward is ended, and thus the effect may be masked.

The present experiment did not investigate skill acquisition directly, but was designed to minimize such training. Subjects were not differentially reinforced on the basis of quality of responses, but on the basis of quantity. Thus there is little likelihood that the Increase and Decrease groups were trained to be more skilful.

Moreover, the fact that reinforcement did not cause the Decrease group to make more responses than the Control group minimizes the possibility that differences in the amount of practice at response writing led to differences in skill level between those groups.

Even though there is no evidence for increased skill with reinforcement, if the reinforced groups came to perceive themselves as more skilful as a result of reinforcement, this could mask the effect. It is noteworthy that the task is a highly practiced response among university students, that of transcribing a brief idea from a book to a piece of paper. It could be argued that university students would perceive themselves as quite highly skilled at this, and that receiving points for handing in responses of a specified quantity but unspecified quality should not reasonably be expected to heighten this perception appreciably.

The question of skill training and reinforcement is an important issue in this area of research, and is dealt with later in the paper. For the present it is sufficient to point out that there is no evidence that students came to view themselves as more competent as a result of the reinforcement manipulation.

c) Reward Salience. Greene and Lepper (1975) point out that for reward to lead to the Overjustification Effect that reward must be salient to the subject. This assertion is backed empirically by a study by Ross (1975) in which it

was demonstrated that salience is a critical variable in showing the effect. Unfortunately, the conditions under which a reward's salience may be determined are not specified by Greene and Lepper. Ross (1975) manipulated the visibility of the reward in his investigation of salience.

In the present case the rewards used seem clearly salient to the subjects. Institution of points contingent upon responses caused a dramatic increase in that response, and discontinuation of this system led to an equally dramatic decrease in the response. Moreover, the rewards were visible to subjects, being reported on their test papers, documents which receive more than passing attention from students. Together these facts can be taken as supporting the conclusion that the reinforcers were visible to the students, that they were in fact reinforcing (in that they changed behaviour), and that the contingencies were unambiguous (students knew which responses were being reinforced).

Thus it is concluded that the failure to demonstrate the Overjustification Effect was not due to the fact that non-salient rewards were used.

d) Initial Interest. It has been pointed out that for interest or responding to decrease, subjects must enter the experiment with some initial response rate at or interest in the task prior to the onset of reward (Greene and Lepper, 1975; Lepper and Greene, 1976). In the present study the mean pre-reinforcement rate of response was 4.48 or 4.96

responses per session, depending upon the treatment of absentees. Only three subjects failed to hand in any responses during Baseline 1 (two in the Control group and one in the Decrease group) and these subjects also failed to make any responses during Baseline 2. Elimination of these subjects does not change the ordering of the results, nor does it affect significance. Thus almost all subjects showed an initial response rate prior to reinforcement, and the elimination of those who did not does not change the conclusions of the study.

It might be argued that subjects were not demonstrating interest in Baseline 1, since they were asked to hand in responses to cues. However, it must be remembered that subjects were instructed that one response would be sufficient but that they could hand in up to 10. The fact that students averaged far in excess of one response per session indicates that the instruction was not the prime factor in their handing in responses, and indicates that the number of responses was relatively free to vary with interest.

Greene (1974) used a similar format when he used a set of arithmetic games which would not have been spontaneously used by pupils. In that case he asked them to use them, and considered the relative response rates to be indicative of interest in the task.

It is argued here that the fact that students engaged in the activity of responding to cues at a rate greater than the required rate, in the absence of any stated constraints,

indicates that the task was initially interesting in the same way that the tasks used by Greene (1974) were interesting.

Summary

It is concluded that there is no evidence that the failure to observe the Overjustification Effect in this study was due to a failure to meet the specified conditions necessary for its occurrence.

Two hypotheses were set out in the introduction to this study. The one which predicted increased interest was not supported by the data. The one which remains is that interest is not decreased by reward.

Accepting a null result as indication of the absence of an effect is scientifically tenuous. Yet the extention of the Overjustification Hypothesis to token economies, and other applied reinforcement systems is extrapolation beyond data presently available. Thus there is only a theoretical basis for making the extrapolation, which the present fair test of the extension did not support, despite the fact that the study was designed to facilitate the possibility of finding support. Until such time as the effect is demonstrated in the context of an ongoing reinforcement system, theoretical grounds for possible deleterious effects of token systems need not be of major concern to practitioners.

The present study does not disprove the Overjustification hypothesis, but provides data indicating some of the limitations of the hypothesis.

GENERAL DISCUSSION

The warning has been made that the use of token reinforcement systems in classrooms may "turn play into work" (Greene and Lepper, 1974) and "interest into druggery" (Lepper and Greene, 1975). The present paper has looked at the evidence upon which these warnings are based. The question being addressed here is highly practical. Do the theories as they have been spelled out have enough empirical support and enough internal consistency to allow general statements about the dangers of reinforcement systems in general?

The Present Studies

These studies were conducted to investigate the question directly, taking predictions derived from the Overjustification Hypothesis, and testing them in the situation in question, an ongoing reinforcement system. Both studies were designed to conform to the conditions described as necessary for the effect to be observed. In addition both studies employed empirically sound reinforcement procedures.

No evidence was found in either study to support the Overjustification Hypothesis. On the contrary, there is evidence in Experiment 1 which indicates that responses to the task (interest) <u>increased</u> because of reinforcement. Several possible explanations for this failure to show the Overjustification Effect were discussed, and did not seem strong enough to account for the obtained lack of effect.

If the Overjustification Hypothesis were a well researched area, backed by a good, empirically derived statement of conditions both necessary and sufficient for detection of the Overjustification Effect, the present studies might easily be dismissed as anomalous. However, at the present state of definition of the hypothesis, sufficient conditions are not described, and only certain possibly necessary conditions are given:

> . . . the use of overly powerful reward systems as predicted by the overjustification hypothesis, may lead individuals with an initial interest in that activity to reattribute their interest in that activity to salient external factors, decreasing subsequent intrinsic motivation . . . we believe that the data from both laboratory demonstrations of overjustification and applied reward programs are sufficient to suggest that reward systems will sometimes simultaneously increase extrinsic motivation and decrease intrinsic motivation (or, alternatively, . . . decreasing the probability of that [reinforced] response in settings where the [appropriate discriminative] stimuli are not present).
>
> (Lepper and Greene, 1976; p. 33 - 34; italics added)

Moreover, empirical support from ongoing token reinforcement systems is lacking.[5] Thus the present studies fill a

conspicuous gap in the current research on the hypothesis.

Logic of Scientific Proof

The Overjustification Hypothesis has been presented and there is a body of research which supports it (see above) Predictions have been made that the Effect extends to applied reinforcement systems, although there is no direct evidence for this. The present results are essentially null results in terms of the hypothesis and its extension, since they fail to show the predicted effect.

Data can be used to refute the Overjustification Hypothesis two ways: 1. By obtaining null results in many situations where the Hypothesis predicted rejection of the Null Hypothesis. By such studies one might show that the hypothesis doesn't hold for a large set of significant circumstances. 2. By finding a pattern of significant results, that are incompatible with any reasonable version of the Overjustification Hypothesis, such a pattern then might best be explained by some different and perhaps unrelated or even contradictory hypothesis. Such an alternative hypothesis would be that under an important set of circumstances, interest is increased by properly applied multiple delivery of reinforcement. (This hypothesis is suggested by the work of Reiss et al. (1975) and Feingold et al. (1975) both of which showed increased responding [interest] following administration of several reinforcers.)

Experiment 1 offers data of the second type contrary to the Overjustification Hypothesis, and supports the

alternative suggested by the work of Reiss et al. and Feingold et al. Experiment 2 offers data of the first type contrary to the Overjustification Hypothesis. Since there is no direct evidence that the Effect occurs in ongoing reinforcement systems, Experiment 2 suggests that the Effect may be limited to conditions other than those existing in properly conducted reinforcement systems. Unless further research should prove it false, Experiment 2 offers strong evidence against an extension of the Overjustification Hypothesis to token economies.

Both experiments are taken as supporting the conclusion that the Overjustification Hypothesis does not apply in properly functioning ongoing reinforcement systems.

The Previous Research

The research which supports the Overjustification Hypothesis is for the most part derived from situations not typical of token economies. In most cases subjects were given one reward only, and the reinforcing value of this reward was not typically assessed. In the one study applying ongoing reward which is taken as supporting the hypothesis (Greene, 1974) reinforcement is confounded with practice effects, and the results cannot be clearly attributed to Overjustification, but are possibly due to fatigue, differences in task difficulty, or satiation. Moreover, Reiss et al. (1975) and Feingold et al. (1975) in studies using multiple reinforcement delivery, demonstrated increased rather than

decreased responding (interest) in the task following the termination of reinforcement. Applied research provides no evidence to support the Overjustification Hypothesis.

There is a large body of literature on token economies, which might seem to be related to this question. However, Greene and Lepper (1975) have presented good arguments for discounting the relevance of such studies to the Overjustification Hypothesis. These studies were not typically designed to test the hypothesis, they dealt with low frequency (interest) behaviour, and may have masked the Effect by teaching skills.

Overall, while previous research may address certain theoretical questions, it has little to say about the present question. The warnings about possible deleterious effects of reinforcement are extrapolations of a theory based on analogue studies, and thus the caveats greatly exceed available data.

How Overjustification Relates To Token Economies

Lepper and Greene (1976) have pointed out that for the Overjustification Effect to be observed it is necessary that reward not be dispensed for low frequency responses, not teach any skill which may lead the subject to sample the intrinsic merits of the response, and not provide feedback about competence.

It has been pointed out by Bornstein and Hamilton (1975) that reinforcement systems are generally instituted in cases where subjects emit a target response at a low

rate or with low accuracy, in the hope of increasing this rate or skill. In addition, since one basic operation of reinforcement is feedback, when reinforcement is contingent upon increased skill, it is likely to give the subject a perception of increased competence. Since these features of reinforcement systems are contrary to those conditions necessary for the Effect to be observed, it is not surprising that it has not been observed in the applied literature.

Nevertheless Lepper and Greene (1976) assert that the Effect may be occurring in all these studies, but, for various reasons detailed above, it is not being detected. Clearly the present results, obtained in studies designed to maximize the likelihood of detecting the Overjustification phenomenon, do not support such an assertion, and bring into question the warning of adverse effects of tokens on interest.

Alternative Theoretical Stances

Reiss and Sushinsky (1975a, 1976) have proposed a Competing Response hypothesis to explain discrepancies between their findings and those of Lepper and Greene, and this hypothesis may have bearing on the present question. They assert that Overjustification may be related to the fact that only one reward is delivered, and feel that the effect dissipates with multiple reinforcement. They have, unfortunately, no strong empirical basis for this view (Lepper and Greene, 1976).

Kruglanski (1975 a, b) has set out a model of endogenous and exogenous motivation, in which he asserts that the location of the intrinsic - extrinsic distinction is not at the subject but at the reward. Some rewards may be seen as task endogenous (money from slot machines) while others seem task exogenous (money for playing checkers).

Kruglanski, Riter, Amitai, Margolin, Shabtai and

Zaksh (1975) conducted two studies testing this hypothesis. In both cases subjects were rewarded for participating in an activity, but in some cases the reward (money) was appropriate (endogenous) to the activity (coin toss game, stock market game) while in other instances the money was task exogenous (playing with blocks, athletics game). In both experiments the Overjustification Effect was observed when money was task exogenous, but not when money was task endogenous (appropriate). In the latter case interest reported by subjects increased with reward, a finding predicted by Kruglanski, but not by Lepper and Greene.

Unfortunately Kruglanski's theory does not specify means of determining a reinforcer's endogenous-exogenous status, except in terms of perceptions of the subjects, and thus is subject to the same criticisms which are levelled against Lepper and Greene.

Thus the alternative theoretical stances, while showing promise for research, offer little in the way of empirical evidence which relates to the present question.

Conclusions

In light of the lack of empirical support for the extrapolation of the Overjustification Hypothesis to applied reinforcement systems; and in light of the fact that several studies, including Experiment 1, show a contrary result; and in light of the fact that Experiment 2 failed to show any indication of the Overjustification Effect, in what would seem to be a fair test of the

hypothesis: it is concluded that there is no basis for concluding that token economies harm interest in any way. It would seem that the burden of proof is upon proponents of the Hypothesis to provide direct evidence of harmful side effects before the issue need be of great concern to practitioners.

FOOTNOTES

1. In response to criticism on this point, Deci, Cascio and Krusell (1975) re-examined the original data from both studies to determine whether any differences existed in the performance of the various groups during training. They assert that no significant differences were found, but unfortunately do not name the analysis used, nor do they report the level of significance the differences failed to achieve. For this reason their response to the original criticism cannot be considered conclusive.

2. Item: Newsletter: The council for exceptional children, Ontario provincial foundation. March, 1975, volume 7, p. 4.

3. This unconservative test of significance was applied in order to maximize the possibility of detecting a difference. This is permissible since this was not a test of the experimental hypothesis, but of a rival hypothesis.

4. Procedures during Training segments were designed to minimize the effect of reinforcement upon responding in those segments, and so relatively simple tasks were used, and the number of responses possible was constrained procedurally. This allowed for control of practice effects (an important omission in many studies in this area) but prohibited a direct assessment of reinforcer effectiveness. Instead an indirect assessment was made using the marble task. This indirect assessment was considered acceptable for two reasons. First the reinforcement procedures were not novel (e.g. Bucher and Davidson, 1974) and were shown to be effective with the present subjects in the same setting with task material of similar difficulty. Secondly, a direct assessment was seen as neither necessary nor sufficient to permit inferences about reward salience. Lepper and Greene (1976) argue that salience is independent of reinforcer effectiveness and should be inferred from the procedures used, since within the Overjustification hypothesis it is possible for a reinforcer which is controlling a subject's behaviour to lack salience as a reward for that behaviour if the subject fails to cognitively view himself as controlled by it.

5. Greene (1974) conducted a study of the Overjustification Hypothesis in a classroom setting using a reasonable token economy. However, as has been pointed out above, the design confounds reward and practice, so that the effect cannot be attributed to one or the other.

6. Subject 13 in the Decrease group made 80 responses in the first session of Baseline 1. This constituted responding to all the cues on the cue sheet, rather than to a small subset as subjects were instructed. It seemed that this extreme score represented a misunderstanding of the instructions, and thus data for this subject in that single session was ignored, while data from remaining sessions was used in calculating means for this subject (see appendix 3).

Subject	Baseln 1	Reinfrce	Baseln 2	Baseln 1	Reinfrce	Baseln 2
1	2.33	5.00	4.00	2.33	3.75	[illegible]
2	4.67	5.25	2.00	4.67	5.25	1.00
3	7.50	6.00	0.00	5.00	3.00	0.00
4	4.00	4.50	7.67	4.00	4.50	5.75
5	7.00	4.50	4.75	7.00	4.50	4.75
6	0.00	0.00	0.00	0.00	0.00	0.00
7	3.00	3.00	2.50	3.33	3.75	1.25
8	1.67	3.25	4.25	1.67	3.25	4.25
9	0.67	3.75	1.00	0.67	3.75	1.00
10	6.00	6.75	5.50	6.00	6.75	5.50
11	5.00	5.50	7.25	5.00	5.50	7.25
12	1.67	2.67	2.50	1.67	2.00	1.25
13	2.50	1.67	0.00	1.67	1.25	0.00
14	5.00	4.50	5.00	5.00	2.25	1.25
15	0.00	3.00	0.00	0.00	3.00	0.00
16	10.00	6.75	6.00	10.00	6.75	6.50
17	5.00	5.00	5.75	5.00	6.75	5.75
18	10.00	10.00	5.00	10.00	10.00	5.00
19	1.67	2.33	3.67	1.67	1.75	2.75

Ignoring Absentees

Absentees equal zero

Mean number of responses handed in by Control group in each experimental condition calculated ignoring absentees and by counting absentees as zero.

Appendix 1.

Subject	Baseln 1	Reinfrce	Baseln 2
1	5.33	7.00	5.25
2	5.33	8.00	8.50
3	2.67	7.00	2.00
4	5.33	7.00	5.25
5	10.00	10.00	8.00
6	2.67	7.50	0.25
7	4.67	7.00	4.25
8	2.50	7.00	2.33
9	3.00	7.00	1.75
10	2.67	6.75	2.75
11	5.00	5.00	1.75
12	3.33	7.00	4.00
13	7.00	7.25	1.75
14	3.00	7.00	0.75
15	3.00	7.00	4.00
16	2.00	7.00	2.25

Baseln 1	Reinfrce	Baseln 2
5.33	7.00	5.25
5.33	8.00	8.50
2.67	7.00	2.00
5.33	7.00	5.25
10.00	10.00	8.00
2.67	7.50	0.25
4.67	7.00	4.25
2.33	7.00	1.75
5.00	7.00	2.75
3.67	6.75	1.75
5.00	5.00	4.00
3.33	7.00	1.75
7.00	7.25	0.75
3.00	7.00	4.00
3.00	7.00	2.25
2.00	7.00	1.75

Ignoring Absentees

Absentees Equal Zero

Mean number of responses handed in by Increase Group in each experimental condition calculated ignoring absentees and by counting absentees as zero.

Appendix 2.

NOT LEGIBLE

Subject	Baseln 1	Reinfrcs	Baseln 2	Baseln 1	Reinfrcs	Baseln 2
	Ignoring absentees			Absentees equal zero		
1	11.33	7.25	6.26	11.33	7.25	6.25
2	[illegible]	7.25	5.50	[illegible]	7.25	[illegible]
3	5.00	[illegible]	[illegible]	[illegible]	5.00	7.25
4	7.00	[illegible]	4.75	7.00	4.25	4.75
5	4.33	5.25	[illegible]	4.33	5.25	0.00
6	2.00	5.25	[illegible]	[illegible]	[illegible]	[illegible]
7	0.00	1.50	[illegible]	0.00	1.00	0.00
8	5.33	5.75	5.00	5.33	5.75	5.00
9	5.33	4.25	[illegible]	5.33	4.25	2.25
10	7.67	7.75	9.67	7.67	7.75	5.00
11	7.33	10.00	5.25	7.33	10.00	5.25
12	2.00	4.25	4.75	1.33	4.25	4.75
13	4.50*	[illegible]	1.75	4.50 *	5.25	1.75
14	3.33	5.00	4.25	3.33	3.75	4.25
15	1.33	[illegible]	2.33	1.33	5.25	2.50
16	[illegible]	4.75	3.33	[illegible]	4.75	[illegible]
17	3.00	[illegible]	1.00	3.00	4.00	0.75
18	6.00	7.00	3.67	6.00	7.00	2.75
19	3.66	4.00	6.75	3.66	4.00	6.75
20	5.50	3.33	3.75	3.67	4.00	3.75
21	5.00	6.67	3.50	5.00	5.00	2.75

Mean number of responses handed in by Increase Group in each experimental condition calculated ignoring absentees and by counting absentees as zero.

Appendix 5

* see footnote 5.

Appendix 4

Control Group

Proportion of sessions where subjects handed in one or more response in each experimental condition, calculated ignoring absentees and by counting absentees as zero

Subject	Ignoring Absentees			Absentees Equal Zero		
	Baseln 1	Reinforce	Baseln 2	Baseln 1	Reinfrce	Baseln 2
1	0.67	1.00	1.00	0.67	0.75	0.50
2	1.00	1.00	0.50	1.00	1.00	0.25
3	1.00	1.00	0.00	0.67	0.50	0.00
4	1.00	1.00	1.00	1.00	1.00	0.75
5	1.00	1.00	1.00	1.00	1.00	1.00
6	0.00	0.00	0.00	0.00	0.00	0.00
7	0.50	0.67	0.50	0.33	0.50	0.25
8	0.67	0.75	1.00	0.67	0.75	1.00
9	0.33	1.00	0.67	0.33	1.00	0.50
10	0.67	1.00	0.75	0.67	1.00	0.75
11	1.00	1.00	1.00	1.00	1.00	1.00
12	0.67	1.00	1.00	0.67	0.75	0.50
13	0.50	0.33	0.00	0.33	0.25	0.00
14	1.00	1.00	1.00	1.00	0.50	0.25
15	0.00	0.50	0.00	0.00	0.50	0.00
16	1.00	1.00	1.00	1.00	1.00	0.75
17	1.00	1.00	1.00	1.00	0.75	1.00
18	1.00	1.00	0.50	1.00	1.00	0.50
19	0.67	0.67	1.00	0.67	0.50	0.75

Subject	Baseln1	Reinfrce	Baseln 2	Baseln 1	Reinfrce	Baseln 2
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.67	1.00	1.00	0.67	1.00	1.00
3	1.00	1.00	0.50	1.00	1.00	0.50
4	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00
6	0.67	1.00	0.75	0.67	1.00	0.75
7	1.00	1.00	0.75	1.00	1.00	0.75
8	1.00	1.00	0.67	0.67	1.00	0.50
9	0.67	1.00	1.00	0.67	1.00	1.00
10	0.67	1.00	1.00	0.67	1.00	0.75
11	1.00	1.00	1.00	1.00	0.75	0.75
12	1.00	1.00	1.00	1.00	1.00	0.75
13	1.00	1.00	1.00	1.00	1.00	0.75
14	0.67	1.00	1.00	0.67	1.00	0.75
15	0.67	1.00	1.00	0.67	1.00	0.75
16	1.00	1.00	1.00	0.75	1.00	1.00

Ignoring Absentees

Absentees equal Zero

Proportion of sessions where subjects handed in one or more response in each experimental condition, calculated ignoring absentees and by counting absentees as zero.

Increase Group

Appendix 5

Subject	Baseln 1	Reinfrce	Baseln 2	Baseln 1	Reinfrce	Baseln 2
1	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	0.75	1.00	1.00	0.75
3	1.00	1.00	1.00	0.67	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00
5	0.67	0.75	0.00	0.67	0.75	0.00
6	0.67	1.00	1.00	0.67	1.00	1.00
7	0.00	0.75	0.00	0.00	0.25	0.00
8	0.67	1.00	0.75	0.67	1.00	0.75
9	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	0.75
11	1.00	1.00	1.00	1.00	1.00	1.00
12	0.50	1.00	0.75	0.33	1.00	0.75
13	1.00	1.00	0.50	1.00	0.75	0.50
14	0.67	1.00	1.00	0.67	0.75	1.00
15	0.67	1.00	1.00	0.67	0.75	0.75
16	1.00	1.00	1.00	1.00	1.00	1.00
17	0.67	1.00	0.33	0.67	1.00	0.25
18	0.67	1.00	0.33	0.67	1.00	0.25
19	1.00	1.00	0.75	1.00	1.00	0.75
20	1.00	1.00	1.00	0.67	0.75	1.00
21	0.50	1.00	1.00	0.33	0.75	0.50

Ignoring absentees

Absentees equal zero

Decrease Group

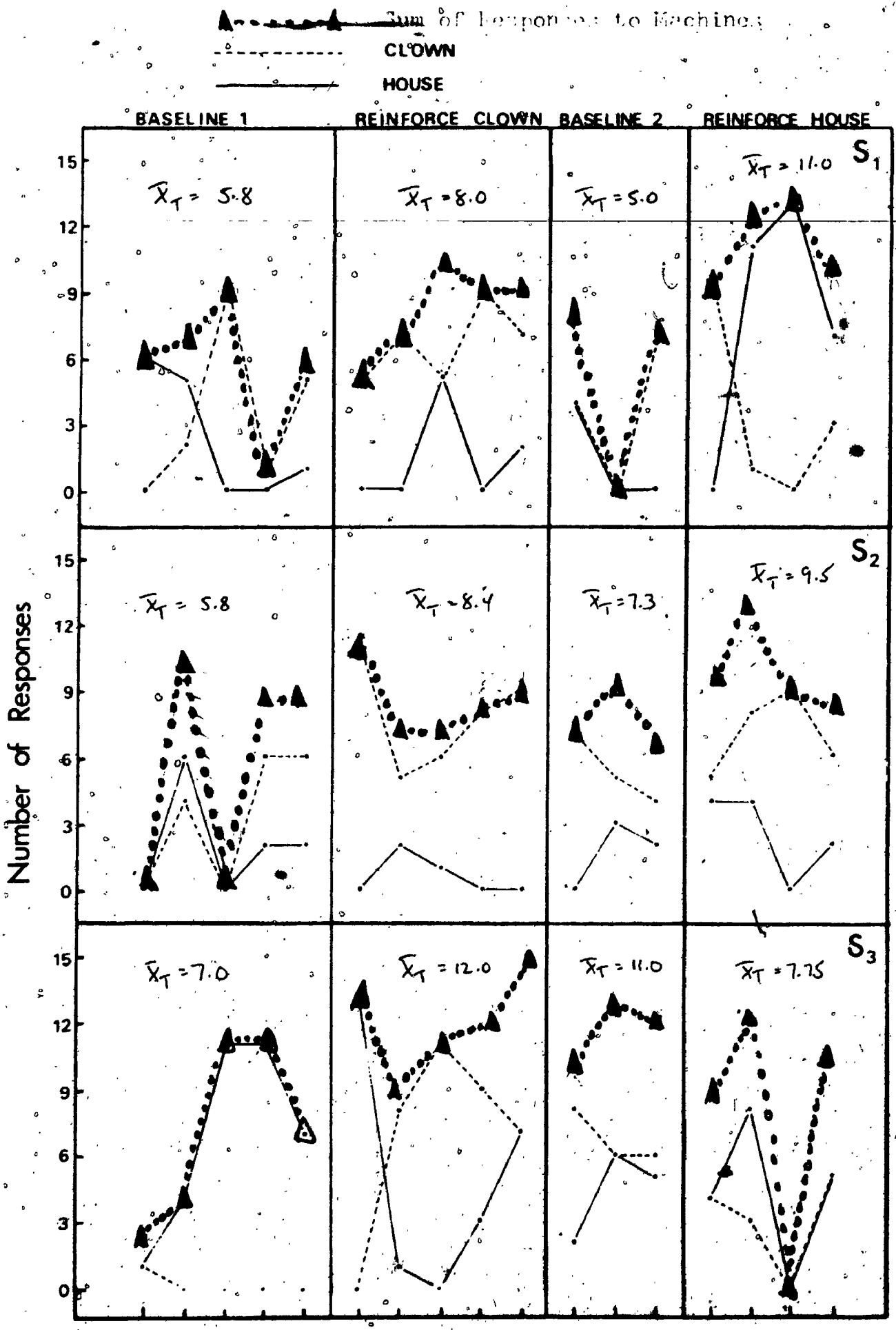
Proportion of sessions where subjects handed in one or more responses in each experimental condition, calculated ignoring absentees and counting absentees as zero

Appendix 6

NOT LEGIBLE

Appendix 7: Sum of Responses to Both Machines. Total responses plotted by session and averaged within conditions for each subject.

NOT LEGIBLE


Sum of Responses to Machines
CLOWN
HOUSE
BASELINE 1
REINFORCE CLOWN
BASELINE 2
REINFORCE HOUSE
Number of Responses
S_1
$\bar{X}_T = 5.8$
$\bar{X}_T = 8.0$
$\bar{X}_T = 5.0$
$\bar{X}_T = 11.0$
S_2
$\bar{X}_T = 5.8$
$\bar{X}_T = 8.4$
$\bar{X}_T = 7.3$
$\bar{X}_T = 9.5$
S_3
$\bar{X}_T = 7.0$
$\bar{X}_T = 12.0$
$\bar{X}_T = 11.0$
$\bar{X}_T = 7.75$
0
3
6
9
12
15

REFERENCES

Bem, D.J. Self-perception theory. In L. Berlowitz (Ed.), Advances in experimental social psychology. Vol. 6. New York: Academic Press, 1972.

Bem, D.J. An experimental analysis of self perception. Journal of Experimental Social Psychology, 1965, 1, 199-218.

Betancourt, F.W. and Zeiler, M.D. The choices and preferences of nursery school children. Journal of Applied Behavior Analysis, 1971, 4, 299-304.

Bornstein, P.H., and Hamilton, S.B. Token rewards and straw men. American Psychologist, 1975, 30, 780-781.

Bucher, B. and Weisz, G. Variables affecting children's preferences for activities. Unpublished manuscript, 1975.

Bucher, B. and Davidson, P. Establishment and generalization of non-compliant responding in children. Canadian Journal of Psychology, 1974, 28, 69-78.

Calder, B.J., and Staw, B.M. The self perception of intrinsic and extrinsic motivation. Journal of Personality and Social Psychology, 1975, 31, 599-605. (a)

Calder, B.J., and Staw, B.M. Interaction of intrinsic and extrinsic motivation: Some methodological notes. Journal of Personality and Social Psychology, 1975, 31, 76-80. (b)

Deci, E.L. Intrinsic motivation, extrinsic reinforcement, and inequity. Journal of Personality and Social Psychology, 1972, 22, 113-120.

Deci, E.L. Effects of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology, 1971, 18, 105-115.

Deci, E.L., Cascio, W.F., and Krusell, J. Cognitive evaluation theory and some comments on the Calder and Staw critique. Journal of Personality and Social Psychology, 1975, 31, 81-85.

DeCharms, R. Personal causation. New York: Academic Press, 1968.

Dickenson, D.J. But what happens when you take the reinforcement away? Psychology in the Schools, 1974, 11, 158-160.

Feingold, B.D., and Mahoney, M.J. Reinforcement effects on intrinsic interest. Behavior Therapy, 1975, 6, 367-377

Gewirtz, J.L., and Baer, D.M. The effect of brief social deprivation on behaviors for a social reinforcer. Journal of Abnormal and Social Psychology, 1958, 56, 49-56.

Greene, D. Comment upon Feingold and Mahoney's "Reinforcement effects on intrinsic interest: Undermining the overjustification hypothesis". Behavior Therapy, 1975, 6, 712-713.

Greene, D. Immediate and subsequent effects of differential reward systems on interest motivation in public school classrooms. Unpublished doctoral dissertation, Stanford University, 1974.

Greene, D. and Lepper, M.R. An information processing approach to intrinsic and extrinsic motivation. Paper presented to the symposium on "The Hidden Cost of Reward" at the annual meeting of the American Psychological Association, Chicago, Illinois, August, 1975.

Greene, D. and Lepper, M.R. Effects of extrinsic rewards on children's subsequent intrinsic interest. Child Development, 1974, 45, 1141-1145. (a)

Greene, D. and Lepper, M.R. How to turn play into work. Psychology Today, 1974, 8(4), 49-54. (b)

Hoppe, R.B. "Token" learning programs. American Psychologist, 1975, 30, 781-782.

Kopel, S. and Arkowitz, H. The role of attribution and self perception in behavior change: Implications for behavior therapy. Genetic Psychology Monographs, 1975, 92, 175-212.

Kruglanski, A.W. An endogenous-attribution theory of intrinsic motivation. Paper presented to the symposium on the "Hidden Costs of Rewards" at the annual meeting of the American Psychological Association, Chicago, Illinois, August 1975. (a)

Kruglanski, A.W. The endogenous-exogenous partition in attribution theory. Psychological Review, 1975, 82, 387-406. (b)

Kruglanski, A.W., Riter, A., Amitai, A., Margolin, B., Shabtai, L., and Zaksh, D. Can money enhance intrinsic motivation? A test of the content-consequence hypothesis. Journal of Personality and Social Psychology, 1975, 31, 744-750.

Kruglanski, A.W., Riter, A., Arazi, D., Agassi, R., Monteqio, J., Peri, I, and Peretz, M. Effects of task-intrinsic rewards upon extrinsic and intrinsic motivation. Journal of Personality and Social Psychology, 1975, 31, 699-705.

Kruglanski, A.W., Alon, S., and Lewis, T. Retrospective misattribution and task enjoyment. Journal of Experimental and Social Psychology, 1972, 8, 493-501.

Lepper, M.R., and Greene, D. On understanding "Overjustification": A reply to Reiss and Sushinsky. Journal of Personality and Social Psychology, 1976, 33, 25-35.

Lepper, M.R., and Greene, D. Turning play into work: Effects of adult surveillance and extrsinic rewards on children's intrinsic motivation. Journal of Personality and Social Psychology, 1975, 31, 479-486.

Lepper, M.R., and Greene, D. Overjustification and intrinsic motivation: Some implications for applied behavior analysis. Paper presented at the annual meeting of the Eastern Psychological Association, Philadelphia, Pennsylvania, April, 1974.

Lepper, M.R., Greene, D. and Nisbett, R.E. Undermining children's intrinsic interest with extrinsic reward: A test of the "Overjustification"hypothesis. Journal of Personality and Social Psychology, 1973, 28, 129-137.

Levine, F.M. and Fasnacht, G. Token rewards may lead to token learning. American Psychologist, 1974, 29, 816-820.

Nisbett, R.E., and Valins, S. Perceiving the causes of one's own behavior. New York: General Learning Press, 1971.

Notz, W.W. Work motivation and the negative effects of extrinsic rewards. American Psychologist, 1975, 31, 884-891.

Parker, R.K., and Nunnally, J.C. Association of neutral objects with rewards: Effects of reward schedules on reward expectancy, verbal evaluation, and selective attention. Journal of Experimental Child Psychology, 1966, 3, 324-332.

Reiss, S., and Sushinsky, L.W. The competing response hypothesis of decreased play effects: A reply to Lepper and Greene. Journal of Personality and Social Psychology, 1976, 33, 233-244.

Reiss, S. and Sushinsky, L.W. Overjustification, competing responses, and the acquisition of intrinsic interest. Journal of Personality and Social Psychology, 1975, 31, 1116-1125. (a)

Reiss, S., and Sushinsky, L.W. Undermining extrinsic interest. American Psychologist, 1975, 30, 782-783. (b)

Ross, M. Salience of reward and intrinsic motivation. Journal of Personality and Social Psychology. 1975, 32, 245-254.

Ross, M., Karinol, R., and Rothstein, M. Reward contingency and intrinsic interest in children: A test of the delay of gratification hypothesis. Journal of Personality and Social Psychology, 1976, 33, 442-447.

Salancik, G.R. Interaction effects of performance and money on self perception of intrinsic motivation. Organizational Behavior and Human Performance, 1975, 13, 339-351.

Skinner, B.F. Science and human behavior. New York: Free Press, 1953.

Walls, R.T., and ~~DiVi~~sta, F.J. Cognitive factors in the conditioning of children's preferences. Developmental Psychology, 1970, 2, 318-324.