DOCUMENT RESUME

ED 365 568 SO 023 151

AUTHOR Berkay, Paul James
TITLE A Critical Analysis of Research on the Overjustification Effect.
PUB DATE 8 May 93
NOTE 51p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Data Analysis; *Data Interpretation; Higher Education; Interpersonal Communication; Researchers; *Research Methodology; Rewards; Social Science Research; State of the Art Reviews; Statistical Analysis
IDENTIFIERS *Overjustification

ABSTRACT

This document examined studies of the overjustification effect. Many studies examining the phenomenon were conducted during the 1970s; findings appeared to be accepted without qualification. It is unclear whether researchers conducted the studies with the proper methodologies and interpreted results correctly. Such factors included the type of reward, expectancy of reward, level of performance demand, and type of feedback. Any study of the overjustification effect should ensure that: (1) claims are properly drawn from data; (2) baseline levels do not differ significantly among treatment groups; (3) valid measures are used for intrinsic interest; (4) results are interpreted properly; (5) only accepted, conventional p values are used; and (6) if claims of behavioral effects of extrinsic rewards are to be made, measures of observed behaviors are used. An analysis of nine such experiments showed that the experiments employed poor methodology and weak or faulty claims. Few studies contained proper claims based on data analysis in both the results and discussion sections. Five studies used unconventional p values. Overall, the articles were not up to the standards of publication in a professional journal. Many qualifications examined in most of the articles are still open to question and await examination in studies with appropriate methodology and design. A reference list identifies the nine studies analyzed; an appendix presents a checklist used in evaluating the studies. (SG)

A Critical Analysis of Research on The Overjustification Effect

Paul James Berkay, M.S.
The University of Oklahoma
Department of Educational Psychology

May 8, 1993

Research findings on the overjustification effect and its qualifications have been cited in recent literature reviews and textbooks (Deci & Ryan, 1985; Reeve, 1992). Many studies on this effect were conducted during the 1970's (Boggiano & Ruble, 1979; Deci, 1971; Dollinger & Thelen, 1978; Lepper, Greene, & Nesbitt, 1973; Rosenfield, Folger, & Adelman, 1980; Sarafino & DiMattia, 1978; Smith & Pittman, 1978). The results from these studies are referenced and appear to be accepted without qualification. One concern might be whether these studies have been conducted with proper methodology and whether the authors have properly interpreted their results. The purpose of this study is to examine some of the early studies of the overjustification effect to determine their merit.

Typical Design Characteristics

In all of the above-referenced studies, there were some common characteristics of most of the designs. Most of the experiments were set up to examine factors that might determine qualifications of the effects of extrinsic rewards on intrinsic motivation. Some of the factors were (a) the type of reward (symbolic vs. tangible), (b) the expectancy of reward (expected vs. unexpected), (c) the level of performance demand (task specific vs. performance specific reward), and (d) the type of feedback (competence information vs. no competence information). Most of the studies focused primarily on one of these factors, although other factors may have been examined. Each study had one or more treatments and typically a control group. In some of the studies, a baseline of intrinsic interest, generally time

spent on the target task, was collected prior to the treatment. Most of the time, subjects were individually brought into an experimental room. In all of the studies, each subject was asked to work on a task (e.g., a puzzle). In the case of an expected reward, subjects would be told about the reward before beginning the task. They would work on the task, usually in the presence of the experimenter. When the treatment period was finished, subjects would receive their award (if that was connected to their treatment condition). A free-choice period followed, in order to determine what effects the treatment had on subsequent intrinsic interest. The free-choice period was generally of one or two varieties. Either the experimenter made an excuse and left the room telling the subject that he/she was free to work on the target task or another task. In other cases, the free-choice period took place up to a few weeks later in the subjects' regular classroom during free-play time. The target activity was left for the subjects, together with other non-target activities. In most cases, during either type of free-choice period, an experimenter would watch the subjects through a one-way mirror and record a measure of intrinsic interest. The most typical measure was amount of time spent on the target task. Treatment group measures would be compared with each other or with their baseline interest. The author would then draw conclusions based on the data related to changes in intrinsic interest from baseline to free-choice period or differences between groups. A

claim of proving or disproving the original overjustification hypothesis with qualifying factors would be made.

Experimental Design Models

Although these studies had several commonalities in their methodology, there were some differences as well. At this point, it might be helpful to present three basic models of design that appeared in most of these studies. One was an ideal model and the other two were far from ideal. First, the ideal model will be presented. This model would allow for the greatest amount of appropriate claims to be made about the experiment. The ideal model is as follows:

Ideal Model

1. A baseline measure of intrinsic interest is taken for all subjects.
2. Subjects are randomized into treatment groups.
3. Through one of many methods (described in a later section), it is determined that there are no significant differences in baseline intrinsic interest between treatment groups.
4. The treatment is administered.
5. The same measure of intrinsic interest taken during the baseline period is taken during a free-choice period.

Three analyses can be conducted with this model as follows:

1. The amount (and direction) of change of intrinsic interest from baseline to free-choice.
2. The difference in amount of change between treatment groups.

3. The difference in measures of free-choice intrinsic interest between treatment groups.

Claims that can be made from these analyses with this model are as follows:

1. For Measure 1 above, for each treatment group, a claim could be made that the treatment resulted in increase/decrease/no change in intrinsic interest.
2. For Measure 2 above, it could be claimed that Treatment A resulted in a greater/lessor/equal change in intrinsic interest when compared to Treatment B.
3. For Measure 3 above, it could be claimed that Treatment A resulted in a higher/lower/equal level of intrinsic interest when compared to Treatment B.

This model is ideal because the maximum amount of claims can be properly made from the results of the data analysis.

Two less ideal models are now presented, the first one is better than the second.

Model 1

1. A baseline measure of intrinsic interest is taken for all subjects.
2. Subjects are randomized into treatment groups.
3. No determination is made about the significant differences between baseline interest among treatment groups.
4. The treatment is administered.
5. The same measure of intrinsic interest taken during the baseline period is taken during a free-choice period.

Three analyses can be conducted with this model as follows:

1. The amount (and direction) of change of intrinsic interest from baseline to free-choice.
2. The difference in amount of change between treatment groups.
3. The difference in measures of free-choice intrinsic interest between treatment groups.

Claims that can be made from these analyses with this model are as follows:

1. For Measure 1 above, for each treatment group, a claim could be made that the treatment resulted in increase/decrease/no change in intrinsic interest.
2. For Measure 2 above, it could be claimed that Treatment A resulted in a greater/lessor/equal change in intrinsic interest when compared to Treatment B.
3. For Measure 3 above, it could be claimed that Group A had a higher/lower/equal level of intrinsic interest when compared to Treatment B.

A minor difference in this model should be noted in Claim 3 above. Differences in levels of intrinsic interest between groups can be determined statistically from Analysis 3, and it can be claimed (due to randomization) that these differences were caused by treatment levels. This claim is not as strong as the one made in the ideal model, however, because the ideal model directly controls for baseline differences, while Model 1 indirectly controls through randomization. Measure 3 taken alone is not of great value because it does not indicate the magnitude

and direction of change in intrinsic interest due to the treatment effects.

Model 2

This is the least desirable model.

1. No baseline measure of intrinsic interest is taken.
2. Subjects are randomized into treatment groups.
3. The treatment is administered.
4. A measure of intrinsic interest is taken during a free-choice period.

Only one analysis can be conducted with this model as follows:

1. The difference in measures of free-choice intrinsic interest between treatment groups.

Only one Claim can be made from this measure as follows:

1. It could be claimed that Group A had a higher/lower/equal level of intrinsic interest when compared to Treatment B.

Similar to Model 2, this measure and claim is acceptable due to randomization, but not as strong as the claim made in the ideal model that directly controls for baseline differences. Again, this measure alone does not give much information about the degree of change resulting from the treatment.

Criteria

In addition to the above models, it might be beneficial to have specific criteria for the evaluation of studies on the overjustification effect. An effective experiment on the overjustification effect should meet the following criteria:

Claims Properly Drawn from the Data

If claims are made about the results of the an experiment and how they prove or disprove the hypothesis, then the following guidelines should be followed:

1. If claims are to be made that a treatment resulted in an increase, decrease, or no change in intrinsic interest, then statistical comparison would need to be made of each treatment group's baseline intrinsic interest with its free-choice period intrinsic interest.
2. If claims are to be made that one treatment resulted in less decrease or increase than another, or that there were no differences, then the amount (and direction) of change from the baseline to the free-choice period for each group would have to be statistically compared. Another option would be to have pretest scores covaried out of post-test scores.
3. If claims are to be made that one treatment resulted in higher, lower, or similar levels of intrinsic interest when compared to another treatment, then subjects should be randomized into treatment groups.

Baseline Level

As mentioned in the last section, baseline levels must not be significantly different between treatment groups. This could be indirectly controlled by randomization. If the experimenter wants to use direct control of the baseline to make stronger claims on differences in free-choice levels of intrinsic interest, the consistency in baseline levels between groups could

be accomplished by (a) using randomized blocks, (b) using a constant level of intrinsic interest in the experiment, or (c) using the baseline level as a variable. For method (a) the following procedure could be followed:

1. After taking baseline measures, subjects could be assigned to blocks based on their initial intrinsic interest levels. The number of subjects in a block should correspond to the number of treatments. Subjects within each block should be randomized to a treatment level. This is similar to the process used to match pairs on a variable in order to perform a correlated t-test.
2. To determine whether the randomization was successful, a statistical analysis could be used to determine that there are no significant differences between treatment groups on baseline measures. An alternative would be to use the pretest scores as covariates.

Method (b) above would be used when the researcher wanted to make claims about the treatment effects on a specific level of initial intrinsic interest. The following procedure could be used for this method:

1. Establish baseline initial intrinsic interest.
2. Determine the level to be used in the experiment and use only those subjects at that level. Eliminate all others.
3. Randomize the subjects to be used into treatment groups.

4. Make sure to statistically compare the baseline interest in the treatment groups in order to establish that it is not significantly different.

An alternative to the above method could be to use the regression approach to ANOVA.

Method (c) could be used to include all levels of baseline intrinsic interest as a variable by using the regression approach to ANOVA. A less preferable option is as follows:

1. Establish baseline initial intrinsic interest.
2. Determine the ranges of interest for each level and then group subjects by these levels.
3. Do a statistical comparison of baseline measures in the different levels to ensure that they are significantly different. These groups should be different, or there is no point in dividing them.
4. Within each level, randomly assign the subjects into treatment groups.

If one of the above methods is followed, then it will be possible to validly compare the free-choice scores from different groups to determine whether there was a significant difference in increase or decrease of scores.

Valid Measures

Valid measures should be used for intrinsic interest. If these measures don't appear to be logically connected to the constructs they purport to be measuring, then the connection should be empirically established.

Proper Interpretation of Results

The results obtained from the experiment should be properly interpreted in the results and discussion sections. Of specific concern is whether results that only approached conventional significance are erroneously claimed to support a hypothesis in the discussion section.

Conventional p values

Claims of significance should be made only on conventional p values acceptable in the Social Sciences. Only results with $p < .05$ should be claimed as significant.

Behavioral vs. Self-reported Measures

If claims of behavioral effects of extrinsic rewards are to be made, then measures of observed behaviors must be used. Self-report of projected behavior from subjects would not be appropriate.

Method

Review Process

In order to evaluate literature on the overjustification effect, a checklist was written to determine how an individual experiment adhered to the criteria proposed above. (See Appendix A.) It might be noted that claims in the results section and in the discussion section were separately examined. It was determined that in some case, the author(s) made faulty claims in one or both sections. Of special concern were the claims made in the discussion section that might be picked up by those browsing the article without examining the data analysis.

A literature review was conducted on the ERIC database, and seven studies that were published in scholarly journals were selected. ERIC documents were not considered because of the lack of a stringent review process. All of the selected articles were written during the 1970's (one was actually dated 1980), as this was the period during which most of the initial overjustification experimentation took place. One of the articles had three experiments (Deci, 1971), while the remaining articles had only one. There were a total of nine experiments reviewed in this study. The purpose of the review was to determine whether the author(s) of each study used correct methodology and drew acceptable conclusions from their data.

The review format was as follows: Each article was summarized and then evaluated by the checklist. Included in the following section are the summaries and checklist evaluations. The experiments are reviewed in chronological order. After review of all articles, an overall checklist summary was made.

Experiment Summaries and Reviews

Experiments 1 - 3

Deci (1971) examined the effects of verbal and monetary extrinsic rewards on intrinsic motivation. Three experiments were conducted. In the first experiment, 24 Introduction to Psychology students were each assigned by class section to one of two treatment groups: monetary reward and no reward. In three separate sessions, subjects were asked to work on puzzles. In the first session, all subjects worked without a reward. In the

second session, the monetary reward group subjects were given $1.00 per completed puzzle and were informed about the reward before commencing work on the task. The control group subjects were given no reward. In the third session, no rewards were given. In each session subjects worked on the puzzles in the presence of the experimenter. Then the subjects were left alone in the room for 8 minutes and told that they could do what they pleased. Subjects could work on the puzzles, read a magazine, or do nothing. The amount of time subjects spent on the puzzles during the 8-minute period was recorded through a one-way mirror. Data were examined to determine whether each group increased or decreased the amount of time spent on puzzles from Time 1 to Time 3. While the monetary group decreased their time, the control group's time increased. The difference in decreased/increased time between the two groups only approached significance, however ($p < .10$). No significant differences were found in this experiment.

A second experiment examined the performance of eight staff members on a college newspaper. Subjects were writing headlines as part of their newspaper assignment. As in Experiment 1, there were two groups: the monetary reward group and the no reward group. During the semester, each group was studied for three sessions. During the first four-week session, all subjects worked for no reward. In the second session, which lasted three weeks, the monetary reward subjects were given 50 cents per headline written, while the no reward group subjects received no

money. In the third session, which was another three-week period, no rewards were given to either group. The experimenter stayed in the newspaper room with the subjects during each session and pretended to be their supervisor. For the measure of intrinsic interest, he recorded the amount of time each subject spent writing each headline. It was assumed that the higher the intrinsic interest, the less time would be spent writing a headline. The absences of each subject were also recorded as a measure of poor attitude. A follow-up session (Time 4) took place five weeks after Time 3. The subjects again wrote headlines without receiving a reward. To examine the effects of the reward on intrinsic motivation, the mean minutes each group spent per headline were analyzed to determine whether motivation increased (lower means) or decreased (higher means) from Time 1 to Time 3. The monetary group had a slight decrease (motivation improvement) in mean time, while the no reward group had a much larger decrease. Although both groups decreased the mean time, there was a significantly greater decrease for the no reward group when compared to the reward group. In examining the increase or decrease in mean time from Time 1 to Time 4, no significant difference was found between the two groups. Although the monetary reward group had a higher percentage of absences for Times 3 and 4 when compared to the no reward group, the differences were not significant.

The third experiment was similar to the first, except that verbal praise was substituted for the monetary reward. The

subjects were 24 Introduction to Psychology students. As in the first experiment there were three sessions. In the first session, all subjects solved puzzles without rewards. In the second session, the verbal reward subjects were given praise, while the no reward group subjects received no performance feedback. In the third session, subjects again worked without reward. Once again, during the 8-minute free period, subjects were left alone and given the choice of working on the puzzles, reading magazines, or sitting and doing nothing. Time spent on the puzzles was observed though a one-way mirror and recorded. The increases or decreases in time spent from Time 1 to Time 3 were examined to determine an increase or decrease in motivation. The verbal reward group slightly decreased in time, while the no reward group greatly decreased the time spent on the puzzles. The no reward group had a significantly greater decrease in time when compared to the verbal reward group. The author claimed that this supported the hypothesis that positive feedback increases intrinsic motivation.

In conclusion, Deci stated that the three experiments demonstrated that monetary rewards decrease intrinsic motivation, while verbal praise increases intrinsic motivation.

Experiment 1 Evaluation

Model. Model 1 (with baseline) was used for this experiment. Subjects were not randomized into groups. They were assigned to a treatment by their class section. It was not determined whether the baseline intrinsic interest was

significant between groups, but the authors looked at difference scores, so this was not a problem.

Analysis. The amount and direction of change from baseline to free-choice and difference in amount of change from baseline to free-choice between groups was analyzed.

Claims in results section. The author made a faulty claim in the results section. He stated that a significant difference with a $p < .10$ supported his hypothesis.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid.

Claims in discussion section. A weak claim was made in the discussion section when the author carried over the faulty claim from the results section stating that a p value of < .10 supported his hypothesis.

Conventional p values. As revealed above, this author made claims on $p < .10$.

Behavioral vs. self-report measures. All claims of behavioral effects of extrinsic rewards in this study were appropriately based on measures of observed behavior.

Experiment 2 Evaluation

Model. Model 1 (with baseline) was used for this experiment. Subjects were not randomized into groups. They were assigned to a treatment by their work shift. It was not determined whether the baseline intrinsic interest levels were significantly different between groups.

Analysis. The amount and direction of change from baseline to free-choice and amount of difference in amount of change from baseline to free-choice between groups was analyzed.

Claims in results section. The author made faulty claims in the result section as follows:

1. Both groups increased in intrinsic motivation from baseline to free-choice period. The control group increased considerably, while the experimental group increased slightly. The author claimed that the experimental group subjects decreased in intrinsic interest.
2. The author claimed that money negatively affected intrinsic motivation, when in fact there was a (likely nonsignificant) increase in intrinsic motivation from baseline to free-choice period for this group.
3. The author claimed that differences with $p < .10$ were significant and supported his hypothesis of an overjustification effect over time and the effect of extrinsic rewards on attitude (as determined by absences).
4. The author claimed that there were differences in absences from baseline to free-choice period for both groups, but he did not check for significance.

Valid measures. The measures used for intrinsic interest in this study appeared to be invalid as follows:

1. The author claimed that faster headline writing meant higher intrinsic interest. It might be possible that the subjects

were bored or apathetic about the task and rushed through it to get it done.

2. The author stated that absences reflected a bad attitude. Other factors, such as personal problems, could have caused absences.

Claims in discussion section. The weak claim of a negative effect of money on intrinsic interest that was made in the results section was carried over to the discussion section.

Conventional p values. As revealed above, this author made claims of significant differences with $p < .10$ on two occasions.

Behavioral vs. self-report measures. All claims of behavioral effects of extrinsic rewards in this study were appropriately based on measures of observed behavior.

Experiment 3 Evaluation

Model. Model 1 was used for this experiment. Subjects were not randomized into groups. They were assigned to a treatment by their class section. It was not determined whether the baseline intrinsic interest levels were significantly different between groups.

Analysis. The amount and direction of change from baseline to free-choice and amount of difference in amount of change from baseline to free-choice between groups was analyzed.

Claims in results section. The author made faulty claims in the result section as follows:

1. The subjects in the experimental groups showed a slight (and unlikely significant) decrease in intrinsic interest from

baseline to free-choice periods. The author claimed that there was no decrease in intrinsic interest without determining statistical significance. (It might be possible that significance was determined, but not reported in the study.) The control group's intrinsic interest mean decreased greatly from baseline to free-choice periods. Again, the author made a claim of this observed decrease without determining significance. Even though these changes appeared to have taken place, the author still should have subjected the observed differences to a statistical analysis.

2. Both groups decreased (or stayed the same) in intrinsic interest from baseline to free-choice period. The author claimed that praise enhanced intrinsic interest.
3. The author used differences with $p < .10$ to support his overjustification hypothesis and to support a claim of performance differences by college major.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid.

Claims in discussion section. Two faulty claims were carried over from the faulty claims made in the results section:

1. Using the $p < .10$ difference to support the overjustification hypothesis.
2. Stating that intrinsic interest was increased by verbal praise when it appeared to have decreased or remained the same.

Conventional p values. As revealed above, this author made claims on $p < .10$ on two occasions.

Behavioral vs. self-report measures. All claims of behavioral effects of extrinsic rewards in this study were appropriately based on measures of observed behavior.

Experiment 4

Lepper, Greene, and Nisbett (1973) examined the effects of extrinsic rewards on the intrinsic motivation of preschool children. A sample of 51 preschool children (ages 4 to 5) from Bing Nursery School at Stanford University were randomized into three groups: expected reward, unexpected reward, and no reward. All subjects were examined in their normal classroom during a baseline period to determine time spent drawing with magic markers. This target activity was one of several activities that the subjects could choose from. Observation was done through a one-way mirror, and the time each child spent drawing was recorded. Two weeks following the baseline period, subjects were brought individually into a room and asked to draw pictures with the magic markers. The subjects in the expected reward condition were told that they would receive a good player award certificate if they drew the pictures. Subjects in the other two groups were not told about a reward. All subjects drew for six minutes. Their drawings were retained so that the quality could be determined by judges at a later date. Upon completion of this session, all subjects in the expected and unexpected reward groups received the certificates. One to two weeks following the

individual sessions, the magic markers and other activities were placed in the classroom during free-play period. The time the subjects spent drawing with the markers was again recorded by the experimenters who were observing through a two-way mirror. It was discovered that in this third session, subjects in the expected reward group spent significantly less time drawing than the students in the unexpected and no reward groups. Increases or decreases from the first to the third sessions were also examined. While intrinsic interest in the unexpected and no reward groups did not significantly change, a significant decrease was discovered with the expected reward group. In addition, the quality of each picture was rated by judges who were blind to the subjects' group membership. It was discovered that the pictures of the expected reward subjects were given significantly lower ratings than the pictures drawn by subjects from the other two groups.

Experiment 4 Evaluation

Model. The Ideal Model was used for this experiment. A baseline was taken for all subjects. To control for differences of baseline intrinsic interest, subject were assigned to blocks and then randomized within blocks to treatment groups. To make these blocks, all subjects with more than four minutes of play during baseline were blocked by their class, blocked by sex within class, then ranked by playing time. There were eight class-sex blocks divided into groups of three. Subjects in each trio were randomized to a treatment condition.

Analysis. The amount and direction of change from baseline to free-choice and differences in measures of free-choice intrinsic interest between treatment groups were analyzed.

Claims in results section. All claims appeared to be properly drawn from the data in the results section.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid.

Claims in discussion section. All claims in the discussion section appeared to be properly drawn from the results of the experiment.

Conventional p values. Only conventional p values ($p < .05$ or less) were used in interpreting significance.

Behavioral vs. self-report measures. All claims of behavioral effects of extrinsic rewards in this study were appropriately based on measures of observed behavior.

Experiment 5

The effects of different types of rewards on intrinsic motivation were examined by Dollinger and Thelen (1978). Sixty preschool and elementary school children (ages 4 to 8) were brought individually into an experimental room and asked to work on mazes. Subjects were randomized into five treatment groups: verbal reward, tangible reward, symbolic reward, self-reward, and no reward. All subjects in the reward treatments were told about the rewards before working on the mazes, so all of the reward conditions used expected rewards. In addition, all rewards were contingent upon successful performance. Subjects in the verbal

reward group received verbal praise when they properly completed a maze. Tangible reward subjects received a pretzel for each successful completion. Subjects in the symbolic condition received a star on a good player award for each successful completion, while subjects in the self-reward condition were told that they could give themselves a star for each successful completion, if they chose to do so. After each subject completed the mazes, the experimenter left the room after explaining that the subject could either work on the mazes or on another activity that was present in the room. During this free-choice period, subjects were observed from behind a one-way mirror. Measures taken were (a) length of time spent on mazes (duration), (b) number of mazes worked on (frequency), and (c) amount of time taken to start working on the first maze (latency). After a data analysis, it was discovered that there were no significant differences in latency time between any of the treatment groups. It was further discovered that subjects in the tangible and self-reward groups spent significantly less time on the mazes and worked on significantly fewer mazes than did subjects in the control group. There were no significant differences between the verbal and symbolic groups and the control group. Subjects in the tangible group worked on significantly fewer mazes than those in the verbal and symbolic groups, while subjects in the self-reward group spent significantly less time on the task than did subjects in the verbal and symbolic groups. The self-reward subjects also completed significantly fewer mazes than did the

symbolic group subjects. There were no significant differences between the verbal and symbolic subjects or between the self-reward and tangible reward subjects. The authors claimed that these results demonstrated that verbal and symbolic rewards are less harmful to intrinsic interest than tangible and self-granted rewards. (The authors claimed that self-reward may have been viewed as extrinsic because they were not given a choice as to whether they would reward themselves or not.)

Experiment 5 Evaluation

Model. Model 2 (no baseline) was used for this experiment. Although subjects were randomized into treatment groups, baseline measures of intrinsic interest were not taken.

Analysis. Only the difference in measures of free-choice intrinsic interest between treatment groups was analyzed.

Claims in results section. Claims are not really made in the results section. The results are merely reported in terms of significance.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid.

Claims in discussion section. In the discussion section, no faulty claims were made.

Conventional p values. Only conventional p values ($p < .05$ or less) were used in claiming significant differences. The authors were careful to report $p < .10$ as marginally significant.

Behavioral vs. self-report measures. All claims of behavioral effects of extrinsic rewards in this study were appropriately based on measures of observed behavior.

Experiment 6

Sarafino and DiMattia (1978) examined the effects of letter grades on intrinsic motivation. A sample of 94 undergraduates in a Personnel Psychology course were administered questionnaires about a fictitious proposed course in Sociology and Casinos. Subjects were led to believe that the course was legitimate. The questionnaire was administered during class time, and each subject was randomly given one of two types of questionnaires representing two grade conditions: letter grades and pass-fail. All subjects received an identical first page that described the course and asked students to rate their interest on an 8-point scale (0 - 7) with 7 as the highest level of interest. Subjects in the grades group had a second page that described a standard grading system (A, B, C, D, and F) for this course, while those in the pass-fail group read a description of a pass-fail grading system. (There was no criterion for passing mentioned.) All subjects were then asked to project their behavior for the proposed class for (a) amount of study, (b) creativity for class work, (c) attendance, and (d) personal satisfaction. They projected these behaviors on an 8-point scale (0 to 7), with 7 indicating the highest degree (e.g. high attendance). For data analysis, each of the four projected behaviors was separately analyzed in a two-way design by projected behavior rating and

initial course interest rating. The only important finding occurred in the analysis of projected study time. When examining all individuals with an initial course interest rating of 7, those in the pass-fail grading condition projected significantly more study time than those individuals in the letter grades group. For subjects with an initial course interest ratings of 6, the results were reversed. Individuals in the letter grade group predicted significantly higher study time than those in the pass-fail group. Individuals with course interest ratings lower than 6, generated projected study times that were not significantly different by grade condition. The authors claimed that these results support the hypothesis that grades negatively effect intrinsic interest of individuals with high initial interest, while they boost the intrinsic interest of those who had low initial interest.

Experiment 6 Evaluation

Model. The Ideal Model was used for this experiment. A report of initial interest, in lieu of a baseline, was taken for all subjects. To control for differences of reported initial intrinsic interest, the reported initial interest was used as a variable. Randomization was accomplished by randomly placing the questionnaires in the stack before they were passed out.

Analysis. The amount of projected interest among initial intrinsic interest groups was analyzed.

Claims in results section. A highly faulty claim was made in the results section. Baseline intrinsic interest was

determined by an 8-point rating scale. In analyzing the study variable, only two initial course interest ratings (6 and 7) corresponded to significantly different study time projections by grading condition. Those with a 7 rating in the pass-fail group showed a higher prediction of study time than those in the letter grades group. This was reversed for those with a 6 rating. The authors appeared to be using this reversal to support their argument that lower initial interest will be increased by letter grades. They are implying that 7 reflects high course interest, while 6 reflects low course interest. It could be argued that both 7 and 6 are high course interest ratings, and that there is little difference between those two points on an 8-point scale. Furthermore, as these two ratings both reflected high initial course interest, they should have produced similar results. The fact that there were no significant differences between grading conditions for those individuals with baseline course interests lower than 6 serves to disprove the authors' claim of letter grades enhancing low intrinsic interest. Those with low baseline course interest in the grades group should have shown significantly higher study projections than those subjects in the pass-fail group, if indeed grades improved intrinsic interest of those with low initial interest.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid, but not as valid as measures of observed behavior.

Claims in discussion section. The same faulty claim of an interaction between baseline level of interest and grade condition stated in the results section was carried through to the discussion section.

Conventional p values. Only conventional p values ($p < .05$ or less) were used in interpreting significance.

Behavioral vs. self-report measures. Claims of behavioral effects of extrinsic interest rewards were all based on measures of self-reported behaviors (for a fictitious course). All measures were projected by the subjects on a questionnaire.

Experiment 7

Two theories accounting for the negative effects of extrinsic rewards on intrinsic motivation were examined by Smith and Pittman (1978). In the distraction theory (Reiss & Sushinsky, 1975, cited in Smith & Pittman, 1978), it was suggested that an extrinsic reward provides distractions (e.g., anticipation) that interfere with enjoyment of a task during a reward period and diminishes later intrinsic motivation during a free-choice period. It is further claimed by this theory that over a large number of extrinsically rewarded trials, subjects will learn to tune out distractions resulting from expected rewards. As a result, they will learn to find the task enjoyable during the reward period and their intrinsic motivation shown during a later free-choice period will not be diminished. In attribution theory, a conflicting theory suggested by Deci (1971), it was explained that a decrease in intrinsic motivation

was caused by the self-perception that a task was performed only to obtain an extrinsic reward. Although not stated in attribution theory, Smith and Pittman suggested that this theory implies that a detrimental effect from an extrinsic reward would remain strong, even after a large number of extrinsically reinforced trials. Both hypotheses were tested with 132 undergraduate students from an Introduction to Psychology course. All subjects were individually brought into a room and asked to solve the Labyrinth skill game. The subjects were randomized into 12 treatment groups based on two factors: reward/distraction and number of trials. For the number of trials variable, to look at the long-term effects of rewards and distractions, three groups were determined. Subjects either performed 10, 25, or 50 trials, and these groups were named the short, long, and medium participation groups respectively. For the reward/distraction variable, there were four conditions. A control group received no rewards and was not distracted. The distraction group was instructed to pay attention to an audio taped lecture while solving all of the puzzles. This group received no rewards. The Reward 1 group received expected monetary rewards. The amount of money per solution remained constant for all three participation groups under the Reward 1 condition. The authors were concerned that there might be a confound caused by those with more hours of participation receiving greater total awards across all trials. To control for this possible confound, in the Reward 2 condition, the short

participation subjects received more money per solution than did the medium participation subjects, while the medium participation subjects received more than the high participation subjects. This allowed all subjects in the three Reward 2 conditions to qualify for the same overall amount of money across all trials. Later analysis determined that there were no confounding effects by overall amount of reward. In subsequent analysis, the Reward 1 and 2 conditions were collapsed. This left a 3 x 3 factorial design with three levels of participation and three reward/distraction levels (reward, distraction, and control). As previously stated, each subject was individually asked to solve the puzzle for the specified number of trials. Those in the reward conditions were told that they could receive money for the solutions, while those in the distraction condition were told they would need to attend to the audio tape while performing their tasks. Following the completion of the prescribed number of trials, each subject was left alone in the room without instructions. The number of trials initiated by each subject during this free-choice period was recorded by a video camera hidden in the ceiling. It was discovered that the control and distraction groups performed significantly more trials during the free-choice period than did the reward subjects. There were no significant differences in the amount of initiated trials between the control and distraction groups. There were no significant differences by level of participation and no interactions between the participation and reward/distraction variables. Further

analysis determined that there was no significant decrease in initiated trials for the reward subjects across levels of participation. This demonstrated that the negative effects of an extrinsic reward on intrinsic motivation did not decrease over several extrinsically reinforced trials. It was determined that for the distraction group, the high participation subjects initiated significantly more trials during the free-choice period than did the low-participation distraction subjects. This showed a diminishing effect of the distraction on intrinsic motivation over several extrinsically rewarded trials. The authors claimed that the results of this experiment support the attribution theory, rather than the distraction theory. This suggests that the overjustification effect is caused by the self-perception that the task was performed for the reward and not for intrinsic reasons.

Experiment 7 Evaluation

Model. Model 2 (no baseline) was used for this experiment. Although subjects were randomized into treatment groups, baseline measures of intrinsic interest were not taken.

Analysis. Only the difference in measures of free-choice intrinsic interest between treatment groups was analyzed.

Claims in results section. No faulty claims were made in the results section.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid.

Claims in discussion section. No faulty claims were made in the discussion section.

Conventional p values. A difference with $p < .07$ was used to support a claim.

Behavioral vs. self-report measures. All claims of behavioral effects of extrinsic rewards in this study were appropriately based on measures of observed behavior.

Experiment 8

The mitigating effects of competence information on the negative effects of rewards on intrinsic interest were examined by Boggiano and Ruble (1979). The subjects were 147 children from two nursery schools (ages 3 to 6) and elementary schools (grades 3 to 5). Elementary and nursery school performance was examined separately. All subjects were individually given a hidden picture task to complete. After completing the task, each subject was left alone after being given the choice of working further on the hidden pictures or another activity. Two independent factors were examined: performance vs. task contingent and positive vs. negative vs. no comparison feedback. These factors resulted in six treatment groups. A seventh group, the control group, was also included. (All subjects, except for those in the control group, received a reward.) Subjects in the performance contingent groups were given candy for meeting specific performance standards, while those in the task contingent groups were given candy for merely completing the task. Subjects in all reward conditions were told ahead of time

about the rewards, so all of the rewards used in this experiment were expected rewards. Upon completion of performance during the treatment period, subjects in the positive comparison groups were told that they performed better than other subjects, while subjects in the negative comparison groups were told that they did worse than others. Subjects in the no comparison groups were given no performance feedback. For both the elementary and nursery school subjects, a control treatment was used that included no reward and no comparative performance feedback. During the free-choice period, the proportion of time each subject spent on the hidden picture task was recorded by an observer watching through a one-way mirror. The data were analyzed, and it was discovered that, for the nursery school subjects, those in the task contingent groups spent significantly less time on the task than did those in the performance contingent groups. There were no significant differences between subjects in the comparison treatment groups. The elementary school children showed different results. The comparative information factor supplanted the contingency factor in affecting intrinsic motivation as measured by time spent on the task during the free-choice period. The positive comparison subjects spent significantly more time on the task than those subjects who were given negative or no comparison information. There was no significant difference on time spent on the task between subjects in the performance and task contingent groups. For the elementary school students, it was determined that a

significantly less amount of time was spent by the task-contingent/no comparison group when compared with the control group (no reward/no comparison). It was also discovered that the negative comparison subjects spent significantly less time on the task during free-play than did the control group. The authors claimed that, based on this experiment, rewards given on the basis of absolute performance standards (performance contingent) without comparative information do not adversely affect intrinsic motivation. For the preschool children, feedback that indicates a high level of performance (performance contingent) sustains intrinsic information when comparative information is not given. For the older children, the highest amount of intrinsic interest is sustained when the reward is based on performance standards and positive comparative information is given. For this older group, rewards given for mere task completion only adversely affect intrinsic motivation when no comparative information is given. It appears that for the elementary school children, comparative information supplants contingency in effecting intrinsic rewards.

Experiment 8 Evaluation

Model. Model 2 was used for this experiment. Baseline measures of intrinsic interest were not taken. The authors did not state whether subjects were randomized into treatment groups.

Analysis. Only the difference in measures of free-choice intrinsic interest between treatment groups was analyzed.

Claims in results section. Several faulty claims were made in the results section. On many occasions, the authors stated that treatments affected intrinsic interest during the free-choice periods. Possible confounds from differing baseline scores were not considered.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid.

Claims in discussion section. In the discussion section, similar to the results section, faulty claims were made that the treatments caused changes in intrinsic interest when no control was made for baseline intrinsic interest differences.

Conventional p values. Only conventional p values ($p < .05$ or $p < .01$) were used to make claims of significant differences.

Behavioral vs. self-report measures. All claims of behavioral effects of extrinsic rewards in this study were appropriately based on measures of observed behavior.

Experiment 9

The effects of reward contingency and competence feed-back on intrinsic motivation were examined by Rosenfield, Folger, and Adelman (1980). The authors proposed that competence feedback derived from a reward, rather than the contingency of the reward, would affect intrinsic motivation. Feedback that reflects a high level of competency in an extrinsic reward would result in higher intrinsic motivation than feedback reflecting low competency or the absence of feedback. A sample of 118 female Introduction to Psychology students were asked individually to work on a

crossword game (Ad-lib). All participants received extra credit toward their course grade. There were three factors and eight treatment conditions. Two factors were combined into one variable and involved the manipulation of contingency and feedback. There were four groups on these two factors: contingent/competency feedback, no-pay/competency feedback, contingent/no-feedback, non-contingent/no-feedback. (In the latter group, the pay rate was purported to be determined randomly.) Each of these four groups were divided into two groups, based on a third factor of pay/ability level. One group consisted of those receiving either high pay and/or high ability feedback, and the other included those getting low pay and/or low ability feedback. (It should be noted all subjects were randomly assigned to one of these two conditions. For the contingent/feedback and no-pay/feedback groups, assignment was not made on the basis of performance, although subjects were led to believe it was. Performance was not actually examined or graded.) In the reward groups, all subjects were told about the rewards ahead of time, so all reward conditions were expected reward conditions. Students in the contingent/competency feedback group learned from the experimenter that high-ability subjects would get more money than low-ability subjects for each completed word in the game. Those in the no-pay/competency feedback group were not offered monetary rewards, but were told that feedback would be given based on levels of performance. Individuals in the contingent/no-feedback group were advised that

subjects would have different pay rates for each completed puzzle. The rates were assigned randomly, rather than based on skill. Those in the non-contingent/no-feedback group were told that pay rates would randomly be assigned, and subjects would be paid by the hour for participation. All subjects were allowed to practice and then left alone for 15 minutes during the treatment period. Following the granting of rewards (for those who were to receive them), each subject was left alone and told that he/she should wait for the experimenter to return with some forms. During this free-choice period, subjects were observed through a two-way mirror, and the amount of time spent working on the game was recorded for each subject. Subjects were also given questionnaires with 15-point scales and asked to rate how much they liked the task. In addition, they were given another 15-point scale and asked to indicate their willingness to return and work on the task again strictly for class credit (without monetary rewards). The data were analyzed and the following was discovered:

1. Low-pay/ability subjects in both competency-feedback groups showed significantly less willingness to come back and liked the task significantly less than the high-pay/ability subjects. The two pay/ability groups from the competency-feedback conditions did not spend significantly different amounts of time on the task. Similar comparison of high- and low-pay/ability subjects in the no-feedback conditions yielded no significant differences.

2. When comparing the contingent/competency-feedback groups to the no-pay/competency-feedback groups, there were no significant differences by pay/ability status. Similar findings were obtained when comparing the contingency/no-feedback groups with the non-contingency/no-feedback groups.
3. For a final comparison, the high-pay/ability subjects in both no-feedback conditions were compared to the high-pay/ability subjects in the contingent/competency feedback condition. (In other words, the no-pay/competency feedback condition was eliminated from this analysis.) The high-pay/ability subjects in the contingent/competency feedback condition showed significantly higher willingness to return than the high-pay/ability subjects in the no-feedback conditions. There were no significant differences between the two high groups on time spent on the task during the free-choice period or likability of the task.

From these results, the authors claimed that contingency did not affect intrinsic motivation. They further claimed that rewards with competence information indicating high performance results in higher intrinsic interest than competence information indicating low performance. High competence information also generates higher intrinsic interest than rewards without competence information.

Experiment 9 Evaluation

Model. Model 2 (no baseline) was used for this experiment. Baseline measures of intrinsic interest were not taken. Subjects were randomized into treatment groups.

Analysis. Only the difference in measures of free-choice intrinsic interest between treatment groups was analyzed.

Claims in results section. Several faulty claims were made in the results section as follows:

1. The authors made claims about differences in free-choice intrinsic interest between groups without determining significance. (Although it might be possible that significance was determined and not mentioned in the article.)
2. The authors claimed that non-significant differences ($p < .08$, $p < .12$, and $p < .16$) supported their hypotheses.

Valid measures. The measures used for intrinsic interest in this study appeared to be valid.

Claims in discussion section. In the discussion section, similar to the results section, faulty claims were made that insignificant differences supported the hypotheses.

Conventional p values. As stated above, unconventional p values (.08, .12, and .16) were used to determine significant differences.

Behavioral vs. self-report measures. Claims of behavioral effects of extrinsic rewards were based on measures of self-reported behaviors. Two projected self-report intrinsic measures

were significantly affected by the treatments, while the only observed behavioral measure (time on task) was not significantly affected, as group differences only resulted in $p < .08$ and .16.

Results

Summary of Evaluations

As previously mentioned, a summary of all experiment evaluations was conducted. This summary is included below. It should be noted that the frequencies reflect the number of experiments, not the number of articles or occurrences. As percentages for only 9 cases can be misleading, only frequencies are shown below.

Criteria	Frequency
Model Used	
Used Ideal Model	2
Used Model 1 (with baseline)	3
Used Model 2 (without baseline)	4
Model Characteristics	
A baseline was taken for all subjects.	5
Subjects were randomized into treatment groups.	5
It was determined that there were no significant differences in baseline intrinsic interest between treatment groups. Method used was:	2
1. Randomized blocks	1
2. Constant level of intrinsic interest	0
3. Baseline as a variable	1
Analyses Conducted	
Amount and direction of change from baseline to free-choice	5
The difference in amount of change from baseline to free-choice periods between treatment groups	3
Difference in measures of free-choice intrinsic interest between treatment groups	6
Claims in Result Section	
Claims were properly drawn from the data.	3
Claims were improperly drawn from the data as follows:	6
1. Claims were based on p values greater than $p < .05$.	4

2. Claims were made about directional changes of intrinsic interest from baseline to free-choice contrary to the actual differences in direction.	2
3. Claims were made about differences between groups or from baseline to free-choice period without looking for significance.	3
4. The author considered a one-point difference on the top end of the baseline rating scale as a separation between high and low intrinsic interest.	1

Valid Measures

The measures used for intrinsic interest appeared to be valid.	8
The measures used for intrinsic interest appeared to be invalid.	1

Claims in Discussion Section

The results obtained from the experiment were properly interpreted in the discussion section.	3
The results obtained from the experiment were improperly interpreted in the discussion section as follows:	6
1. Claims of significant differences were based on p values greater than $p < .05$.	3
2. Claims were made about directional changes of intrinsic interest from baseline to free-choice contrary to the actual differences in direction.	2
3. The author considered a one-point difference on the top end of the baseline rating scale as a separation between high and low intrinsic interest.	1

Conventional p Values

Only conventional p values ($p < .05$ or less) were used to claim significant differences.	4
Unconventional p values (ranging from .07 to .16) were used to claim significant differences.	5

Behavioral vs. Self-reported Measures

Claims of behavioral effects of extrinsic rewards were appropriately based on measures of observed behaviors.	7
Claims of behavioral effects of extrinsic rewards were based on measures of self-reported behaviors.	2

Discussion

The above summary shows that most of these experiments contained poor methodology and weak or faulty claims. Only two studies used the ideal model. Sarafino & DiMattia (1978) may have used this model, but their interpretation of their results was among the poorest of all of these studies. Lepper, Greene, and Nesbitt (1973) also used the ideal model and conducted a nearly flawless study. Four experiments used Model 2 (without a baseline), and any results from this model might not be quite as strong as those from the ideal model that directly controlled for initial intrinsic interest.

What may have been most disturbing was that only three studies contained proper claims based on the data analysis in both the results and the discussion sections. In the remainder of the studies, improper claims were made in the results sections

and then carried over to the discussion sections. This can be most troublesome as some individuals will read only the discussion sections and then conclude that the overjustification effect and its qualifications have been supported. Some of these careless individuals have gone on to uncritically reference these findings in textbooks, so that other unsuspecting individuals might also accept the questionable findings without question.

It was also disturbing that five out of the nine experiments contained claims of supported hypotheses based upon differences determined by unconventional p values (greater than $p < .05$).

These studies were not entirely flawed. On the positive side, there were not too many problems with the use of invalid measures of intrinsic interest or claims of behavioral effects of intrinsic interest based on measures of self-reported behavior.

Despite these few positive points, it might be clear that the articles were not up to the standards of publication in a professional journal. One might wonder how these studies were able to pass the stringent review process of a refereed journal with these flaws going undetected.

Even though most of the studies were below standard, it was fortunate that one major study, Lepper et al. (1973), was beyond reproach and did provide strong evidence for the overjustification effects and the negative effects of expected rewards on intrinsic motivation. Many of the other qualifications examined in the other articles are still open to question and await examination in studies with appropriate

methodology and design. It is interesting that Lepper et al. (1973) was one of the earlier studies on overjustification, but the later researchers didn't use the Lepper et al. design as a model for their own endeavors.

It is suggested by this author that the Lepper et al. (1973) design be used as a model for future research in examining qualifications of the overjustification effect.

REFERENCES

Boggiano, A. K., & Ruble, D. N. (1979). Competence and the overjustification effect: A developmental study. Journal of Personality and Social Psychology, 37(9), 1462-1468.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology, 18(1), 105-115.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum.

Dollinger, S. J., & Thelen, M. H. (1978). Overjustification and children's intrinsic motivation: comparative effects of four rewards. Journal of Personality and Social Psychology, 36(11), 1259-1269.

Lepper, M. R., Greene, D., & Nesbitt, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. Journal of Personality and Social Psychology, 28(1), 129-137.

Reeve, Johnmarshall. (1992). Understanding motivation and emotion. Fortworth, TX: Harcourt, Brace, Jovanovich College Publishers.

Rosenfield, D., Folger, R., & Adelman, H. F. (1980). When rewards reflect competence: A qualification of the overjustification effect. Journal of Personality and Social Psychology, 39(3), 368-376.

Sarafino, E. P., & DiMattia, P. A. (1978). Does grading undermine intrinsic interest in a college course? Journal of Educational Psychology, 70(6), 916-921.

Smith, T. W., & Pittman, T. S. (1978). Reward, distraction, and the overjustification effect. Journal of Personality and Social Psychology, 36(5), 565-572.

Appendix A

Checklist

Model Used

____ Ideal Model

____ Model 1 (with baseline)

____ Model 2 (without baseline)

Model Characteristics

___ A baseline was taken for all subjects

___ Subjects were randomized into treatment groups

___ It is determined that there are no significant differences in baseline intrinsic interest between treatment groups. Method used was:

___ randomized blocks

___ constant level

___ baseline as a variable

Analyses Conducted

___ Amount and direction of change from baseline to free-choice intrinsic interest

___ Difference in amount of change from baseline to free-choice intrinsic interest between treatment groups

___ Difference in measures of free-choice intrinsic interest between treatment groups

Claims in Results Section

____ Claims were properly drawn from the data

____ Claims were improperly drawn from the data as follows:

Valid Measures

____ The measures used for intrinsic interest were valid.

____ The measures used for intrinsic interest were invalid as follows:

Claims in the Discussion Section

____ The results obtained from the experiment were properly interpreted in the discussion section.

____ The results obtained from the experiment were improperly interpreted in the discussion section as follows:

Conventional p values

____ Only conventional p values ($p < .05$ or less) were used in determining significant differences.

____ Unconventional p values were used in determining significant differences as follows:

Behavioral vs. Self-Reported Measures

____ Claims of behavioral effects of extrinsic rewards were appropriately based on measures of observed behaviors.

____ Claims of behavioral effects of extrinsic rewards were based on measures of self-reported behaviors as follows:

Author Notes

Paul Berkay received an M.S. degree in Counseling through the Educational Psychology Department at California State University, Long Beach, and is currently pursuing a Ph.D. in Education at The University of Oklahoma.

Inquires should be addressed to the author in care of The University of Oklahoma, Department of Educational Psychology, 820 Van Vleet Oval, Norman, Oklahoma 73019-0260.