

Social Loafing and Group Evaluation

Stephen G. Harkins
Northeastern University

Kate Szymanski
Adelphi University

Recent work (e.g., Harkins, 1987; Szymanski & Harkins, 1987) has suggested that social loafing occurs because participants' outputs cannot be evaluated by the experimenter, by the coactors, or by the participants themselves. This analysis has focused on the output of the individual, but in loafing research, participants work together to produce a group product. However, in this prior work participants have been unable to make anything of this group product, because no standard of comparison has been made available. Several recent formulations (e.g., Goethals & Darley, 1987) have suggested that the potential for group evaluation could motivate performance. Testing this hypothesis in 2 experiments, 1 using an optimizing task and the other a maximizing task, we found that providing a standard that allowed the "group" to evaluate its performance eliminated the loafing effect. The implications of these findings for current theories of group evaluation are discussed.

Kerr (1983) has identified several types of group motivation loss effects (e.g., free-riding, the sucker effect, and social loafing), and has suggested that different processes account for each of these effects. For example, Kerr proposed that free-riding stems from the participant's perception that he or she could benefit from the others' contributions, whereas sucker effects arise from the participant's perception that the others could profit from the participant's contributions (i.e., the others are taking advantage of the participant by free-riding). It has been argued that the social loafing effect is mediated by the participant's perception that his or her individual outputs cannot be evaluated (Harkins, 1987). In this article, we focus on the loafing effect.

Reduced effort, termed *social loafing* (Latané, Williams, & Harkins, 1979), has been obtained when outputs have been combined such that individual contributions could not be isolated. For example, Latané et al. (1979) asked participants to shout as loudly as they could, individually and ostensibly in groups of 2, 4, and 6. When participants "shouted together,"¹ their joint outputs were measured by a single sound-level meter that registered the total output. Other loafing experiments share this pooling feature, whether participants have been asked to pull on a rope (Ingham, Levinger, Graves, & Peckham, 1974), to pump air (Kerr & Bruun, 1981), to generate as many uses as possible for an object (Harkins & Petty, 1982), to watch for seldomly occurring signals (Harkins & Petty, 1982), to react to persuasive messages (Petty, Harkins, & Williams, 1980), to evaluate proposals (Brickner, Harkins, & Ostrom, 1986), or to negotiate simple mazes (Jackson & Williams, 1985). Under these conditions, participants could receive neither credit nor blame for their individual performances, and so they loafed.

Williams, Harkins, and Latané (1981; see also Kerr & Bruun, 1981) showed that when participants were led to believe that their individual outputs could be identified, the loafing effect

was eliminated whether they shouted individually or together. They argued that these results suggested that identifiability was an important mediator of the loafing effect. However, Harkins and Jackson (1985) showed that more than identifiability alone was required to eliminate the loafing effect. Participants were asked to generate as many uses as possible for an object that they were led to believe was the same as or different from the object for which the other members of their group were generating uses. Participants whose outputs were identifiable, but who thought that they were generating uses for different objects, generated as few uses as did those whose outputs were pooled. Harkins and Jackson (1985) argued that these participants loafed because, even though their outputs were identifiable, they thought that these outputs could not be compared with those of the participants whom they thought had generated uses for different objects. Loafing was eliminated only when outputs were identifiable *and* participants believed that they were generating uses for the same object, because it was only then that individual evaluation was possible.

This line of argument suggests that it is the potential for evaluation that mediates the loafing effect, and that for evaluation to be possible, two pieces of information are necessary: some measure of output (e.g., how many uses a person generated) and some sort of standard against which this output can be compared (e.g., the number of uses generated by others in the group). The number of uses generated by others serves as an example of a social standard, and on maximizing tasks (Steiner, 1972), which require the participants to produce as much as possible, this is the only type of standard that is available. On optimizing tasks (Steiner, 1972), which require participants to achieve some criterion performance, both social (e.g., the performance of other participants) and objective (e.g., how closely the participants approached the criterion) standards are avail-

Correspondence concerning this article should be addressed to Stephen G. Harkins, Department of Psychology, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115.

¹ Of course, in the Latané et al. (1979) experiment, participants always shouted individually in the crucial conditions, but were led to believe that they were shouting with others in the various-sized groups.

able.² For example, on a vigilance task in which participants are asked to detect infrequently occurring signals, a social standard would be the average number of signals detected by previous participants, whereas an objective standard would be the number of signals actually presented.

A review of previous loafing research suggests that when outputs have been pooled, one or both of these pieces of information have been missing. For example, in the research in which the shouting paradigm was used (Latané et al., 1979), when participants "shouted together," there was no way to determine what each person's contribution was; thus, the output component was lacking. In the Harkins and Jackson (1985) experiment, when the participants' outputs were identifiable but they thought that they were generating uses for different objects, there was no standard even though the output component was present. The pooled conditions of other loafing experiments are amenable to similar interpretation.

This analysis accords the potential for evaluation a central role in producing loafing effects, but leaves ambiguous which of the potential sources of evaluation available in the paradigm are responsible for producing the effect. Szymanski and Harkins (1987) identified three potential sources of evaluation in the loafing paradigm. Participants may loaf because the experimenter is unable to evaluate their performances, because their coparticipants cannot evaluate them, or because they are unable to evaluate themselves. Harkins, Latané, and Williams (1980) suggested that the experimenter was the critical source of evaluation in the paradigm, which is consistent with previous descriptions of the tasks used in typical loafing research as boring and uninteresting (e.g., Brickner et al., 1986; Harkins et al., 1980). Unless there is an external agent monitoring one's performance, why should one put much effort into such tasks? However, Szymanski and Harkins (1987) showed that although the potential for evaluation by the experimenter alone is certainly sufficient to eliminate the loafing effect, it is not necessary. Providing participants with the potential for self-evaluation was sufficient to eliminate the loafing effect, even on the seemingly dull, uninteresting tasks used in loafing research.

Szymanski and Harkins's (1987) analysis focused on evaluation of the *individual*. That is, it suggested that people loaf because their *individual* outputs cannot be evaluated by the experimenter, coactors, or the self. However, in addition to output at the individual level, there is also the group's output. This output could be compared with a standard, allowing evaluation at the group level.

Several recent formulations have suggested that the potential for group evaluation could motivate performance. For example, Goethals and Darley (1987) have proposed a revision of social comparison theory that incorporates group comparison. They noted that, in the original theory (Festinger, 1954), the notion of self-knowledge was emphasized; however, they argued that the concern for self-validation is also important. Individuals do not want simply to inform themselves about how they stand relative to others; they also want to discover that they stack up well. Of course, there was some suggestion of this in the original theory—in Festinger's (1954) notion of the unidirectional upward drive for abilities—however, in their revised theory, Goethals and Darley proposed that self-validation and self-knowledge should be weighted equally.

In addition to revising the theory by giving equal billing to concerns about self-validation, Goethals and Darley (1987) have proposed to extend the theory by including social comparison at the group level. They begin this extension by incorporating Tajfel and Turner's (1986) social identity theory. Tajfel and Turner (1986) suggested that people strive to maintain a positive self-concept, which can be achieved by attaining a positive social identity. One means of gaining such a positive identity is to discover that one's group compares favorably with other groups. If, on the other hand, one's group fares badly in such comparisons, the individual will attempt to improve the group, or failing that, will leave it. Goethals and Darley saw this concern about positive self-concept as corresponding to the self-validation aspect of their revised social comparison theory, but at the group rather than the individual level.

To this group-level concern with self-validation, Goethals and Darley (1987) have proposed to add a group-level concern with self-knowledge. That is, in addition to a concern about how one's group stacks up against other groups, one could simply be concerned about gaining information about how well one's group does at various tasks, which would allow the group to function more effectively. By making this addition, Goethals and Darley have provided the outline of a social comparison theory that works at the group as well as the individual level. Working from this revised theory of social comparison processes, one could predict that the possibility of group evaluation could motivate performance in the absence of the potential for individual-level evaluation by any source.

However, a review of loafing research suggests that this opportunity has not been available in the previous work, because once again, one or both of the required pieces of information—performance output and a standard—have been missing. For example, in the Latané et al. (1979) research, participants were told that they would be shown their scores at the end of the session. However, there was no standard against which this group output could be compared. In other loafing research, participants have either been told explicitly that they would not be shown their output (e.g., Williams et al., 1981) or there was ambiguity about whether it would be shown (e.g., Harkins & Petty, 1982; Jackson & Williams, 1985). Even if participants thought that their group would receive output feedback, there was no reason for them to believe that any standard would be available against which they could compare their group's performance.

Experiment 1

As suggested by Goethals and Darley (1987), providing participants with the opportunity for group evaluation may improve performance. In our first experiment, we tested this hypothesis by manipulating the participants' opportunity to evaluate their group's performance, while attempting to minimize the salience of the possibility of evaluation of the individual

² On tasks that are at least somewhat familiar, participants could also use some sort of personal standard (e.g., how well I did compared with the last time I did it); however, they are unlikely to have had any prior experience on the tasks used in this research. Thus, such personal standards are not likely to be available.

scores by the experimenter, coparticipants, and self. Otherwise, any improved performance could be attributed to the possibility of evaluation by these sources.

We used a vigilance task in which participants were asked to detect as many signals as they could while minimizing false alarms (see Harkins & Petty, 1982). In the crucial condition, the experimenter told participants in a pooled output condition that, at the end of their session, they would be provided with the number of signals detected by their group and the number of signals that would have been detected had each person in their group detected each signal. This latter piece of information served as an objective standard against which the participants could compare their group's performance on this optimizing task. At the same time, we attempted to minimize the possibility that participants felt that evaluation of their individual outputs by other sources was possible.

Method

Participants. The participants were 144 male and female undergraduates who were run in groups of 3 and who participated as a means of earning extra course credit. The design was a 2 (individual scores vs. pooled) \times 2 (standard vs. no standard) factorial with 12 groups randomly assigned to each of the four cells.

Procedure. On arrival, the participants were seated at a semicircular table with partitions and asked to read the following instructions that were presented on a TV monitor:

We are interested in studying the performance of groups and individuals on vigilance tasks. The vigilance task requires you to watch for a dot to flash on a TV monitor. When you see the dot you are to signal by pressing a button. You will be watching one fourth of the TV screen. In this session, you will be watching the upper left-hand quadrant of the screen. Within this quadrant, the dot may flash in any location since each position is randomly determined by the computer.

Groups in the individual scores conditions then read,

We are interested in your performance in your group. Over the course of the experiment your individual scores will be recorded. If, for example, there were 10 signals in the task, a perfect score for you would be 10, which shows that you reported the signal each time it appeared. Of course, if one of you missed the signal, while the others got it, the people who detected the signal would get credit, while the one who didn't, would not. Please try to detect as many of the signals as you can while minimizing the number of times you falsely report the presence of a signal.

Groups in the pooled conditions read,

We are interested in the performance of your group. Over the course of the experiment your group's score will be represented by the sum of your individual performances. If, for example, there were 10 signals in the task, a perfect score for your group would be 30 which would show that all of you reported the signal each time it appeared. Of course, if one of you missed the signal while the others got it, your group would get credit for only two detections. Please try to detect as many of the signals as you can while minimizing the number of times you falsely report the presence of a signal.

Crossed with this manipulation was the manipulation of standard. Participants in the individual scores-standard condition read:

At the end of the experiment you will be shown how many signals you detected. We will also tell you how many times the dot flashed.

Of course, this information could influence the performance of others, so please do not discuss this experiment with anyone.

In the pooled-standard condition, participants read that their group would be shown how many signals their group had detected *and* the number of signals that would have been detected if each person had detected every signal (the standard). They were also cautioned against discussing their results. Participants in the no standard conditions read either that they would be shown how many signals they (individual scores-no standard) or how many their group (pooled-no standard) had detected, but that they would *not* be shown "how many dots flashed since this information could affect the performance of later participants."

All participants then went through a 1-min practice trial in which three signals flashed. At the end of the practice trial, participants either saw individual feedback (individual scores) or the sum of the number of signals detected by the group and the sum of the number of false alarms (pooled). All participants were then told that there were three signals in the practice trial, and participants in the standard conditions were reminded that they would be shown how many signals had occurred at the end of the session. No standard participants were reminded that they would not be given this information so as to preserve confidentiality.

The experimenter then initiated the task and left the room. Across the 14-min task, 14 signals occurred. These signals were each composed of one graphics block and were flashed for 1/30th of a second at randomly determined locations in the upper left-hand quadrant of the TV screen. At the end of the task, the experimenter asked participants to answer a set of questions meant to gauge the success of the manipulations. They were asked to rate the extent to which the experimenter would be able to tell exactly how many dots they had detected, the extent to which the other participants would be able to tell how many dots they had detected, the extent to which the participants would be able to tell how many dots they themselves had detected, and the extent to which they would be able to tell how well they themselves had performed. Participants in the pooled conditions were also asked to what extent they would be able to tell how well their group performed. Each of these questions was accompanied by an 11-point scale anchored by *Not at all* (1) and *Know exactly* (11).

Results

We analyzed the data in 2 (individual scores vs. pooled) \times 2 (standard vs. no standard) analyses of variance (ANOVAs). We used the group as the unit of analysis, and where appropriate, Tukey's honestly significant difference (Kirk, 1982) as the a posteriori test.

Manipulation checks. Participants in the individual scores conditions felt that the experimenter would be better able to tell how many dots they detected ($M = 10.2$) than did participants in the pooled conditions ($M = 6.3$), $F(1, 44) = 34.2$, $p < .001$. Participants in the individual scores conditions felt that the other participants would be more able to determine how many dots they detected ($M = 8.8$) than did participants in the pooled conditions ($M = 3.0$), $F(1, 44) = 122.2$, $p < .001$. The individual scores participants also felt that they would be better able to tell how many dots they had detected ($M = 9.3$) than did pooled participants ($M = 6.4$), $F(1, 44) = 27.5$, $p < .001$. Finally, participants in the individual scores conditions felt that they would be better able to tell how well they performed ($M = 8.7$) than did participants whose outputs were pooled ($M = 5.7$), $F(1, 44) = 33.8$, $p < .001$. No other effects were reliable. Participants in the pooled-standard condition reported that they would be

better able to tell how well their group performed ($M = 8.5$) than did pooled-no standard participants ($M = 5.2$), $F(1, 22) = 17.3, p < .001$.

Errors. Previous research (Harkins & Petty, 1982) has shown that, on this vigilance task, the two possible types of errors (misses and false alarms) are positively related. For this reason, we added these two types of errors for analysis, which revealed two reliable main effects and an interaction.

Individual scores participants made fewer errors ($M = 3.7$) than did pooled participants ($M = 5.7$), $F(1, 44) = 9.6, p < .001$, and standard participants made fewer errors ($M = 3.7$) than did no standard participants ($M = 5.7$), $F(1, 44) = 9.1, p < .001$. However, these main effects must be interpreted in terms of the reliable interaction, $F(1, 44) = 12.6, p < .001$. Participants in the pooled-standard condition made as few errors ($M = 3.6$) as did participants in the individual scores conditions (individual scores-standard $M = 3.9$, individual scores-no standard $M = 3.5, ps > .20$), and participants in each of these conditions performed better than did participants in the pooled-no standard condition ($M = 7.9, ps < .05$).

Discussion

Analysis of the manipulation checks suggests that we were successful in creating the conditions required for a test of the hypothesis. Participants in the pooled conditions felt that they could be evaluated by the experimenter, their coparticipants, and themselves to a lesser extent than did individual scores participants. In addition, participants in the pooled conditions did not differ in the extent to which they felt that they could be evaluated by any of these sources ($ps > .20$). The only difference between the pooled conditions was that the pooled-standard participants felt that they could evaluate their group to a greater extent than did pooled-no standard participants.

Without a standard, participants whose outputs were pooled made more errors than did individual scores participants, a replication of the loafing effect (e.g., Harkins & Petty, 1982). With a standard, participants whose outputs were pooled performed as well as the individual scores participants. These findings are consistent with Goethals and Darley's (1987) notion that the opportunity for group evaluation is sufficient to enhance performance.

Experiment 2

In our second experiment, we attempted to determine the generalizability of the group evaluation effect by using a different type of task. Instead of using an optimizing task (achieve a criterion), we used a maximizing task (do as much as you can; Steiner, 1972). Participants were asked to generate as many uses as possible for a common object, a knife. In the crucial condition, participants whose outputs were pooled were told that, at the end of their session, they would be provided with the average number of uses generated by all groups tested in a previous version of the experiment. This average served as a social standard against which the participants could compare their group's performance.

In addition to testing the generalizability of the group evaluation effect, we sought to test an assumption that was made in

Experiment 1. To establish the conditions necessary for a test of our hypothesis, we argued that the salience of the evaluation potential of each of the sources should be minimized, but the focus was on the ability of these sources to evaluate *individual* outputs. We made the argument that, in previous loafing research, participants could not evaluate their group's performance because they lacked information about the standard, their output, or both. In Experiment 1, we manipulated the participants' access to a standard and found support for the notion that the opportunity for group evaluation could motivate performance. However, in the earlier loafing research, the experimenter *did* have access to both a standard (e.g., the previous group's performance) and the group's output, and thus could evaluate the group's performance. Yet it was under exactly these conditions that the loafing effect was obtained. This suggests that although the experimenter's ability to evaluate individual performances motivates individual performance, the experimenter's ability to evaluate group performances does not motivate performance at the group level. In Experiment 1, we assumed that, as in previous research, participants in the pooled conditions believed that their group's output could be evaluated by the experimenter, but that they were not motivated by this information.

In Experiment 2, we tested this assumption by including a social loafing replication condition in which no mention was made of a standard, and by asking participants in each of the pooled output conditions to what extent they believed that the experimenter could evaluate their group's performance. To support our interpretation, we expected to find that all participants whose outputs were pooled would report that the experimenter could evaluate their group's performance, whether or not they were provided with a group standard, but that improved performance would be observed only when the participants could evaluate their groups.

The experiment's design was $2 \times 2 + 1$ with one missing cell. Groups of 3 participants were led to believe that their scores would be collected individually or would be combined with those of the other participants. Crossed with this manipulation, groups read either that they would be told the average performance of previous participants (a social standard) or that they would not be given this information. Of course this combination of manipulations yields four conditions, but one of them, the individual scores-no standard condition, was not included. In previous research (e.g., Harkins & Szymanski, 1988; Szymanski & Harkins, 1987), it has been found that the level of performance exhibited in this condition was equivalent to that achieved in the individual scores-standard condition. The primary focus of the present experiment was on the pooled conditions, and we thought that the individual scores-standard condition would provide the point of comparison required to answer the question of interest. Thus, we included only three cells from the basic 2×2 design. We included a social loafing replication condition as the fourth condition.

Method

Participants. The participants were 192 male and female undergraduates who were tested in groups of 3 and who participated as a means of earning extra course credit. The design was 2 (individual scores vs.

pooled) \times 2 (standard vs. no standard) + 1 (social loafing replication) with one missing condition (individual scores–no standard). Groups were randomly assigned to one of the four experimental conditions.

Procedure. Participants were seated at a semicircular table separated by dividers and asked to read the following instructions:

We are interested in the performance of groups and individuals on a task called brainstorming. You will be given the name of an object and your task will be to come up with as many uses for this object as you can. These uses can be ordinary or unusual. Don't be concerned about quality. Simply write down as many uses as you can.

Each of you will be given slips of paper, provided with the name of the object, and told to begin. You should write each use you generate on a slip of paper and then fold this slip and slide it down the tube you see. A bell will sound when you are to stop. If you finish before the bell sounds, simply wait quietly until it does sound.

In front of each of the participants was a tube that extended into a box that was on the floor in front of the table.

Participants then read instructions that placed them into one of the four experimental conditions. Those assigned to the individual scores condition read,

We are interested in the number of uses generated for this object by each of you. So at the end of the experiment we will count the number of uses generated by each of you individually and tell you how many uses each of you generated.

Individual scores participants were then shown that the box was divided into three compartments so that their individual scores could be ascertained.

Participants in the pooled conditions read,

We are interested in the number of uses generated for this object by your group. So at the end of the experiment we will count the total number of uses generated by all of you and tell you what this total is.

In the social loafing replication condition, the experimenter also told participants that we were interested in the number of uses generated by the group and that we would count the total number of uses generated by all of them at the end of the experiment. However, participants were not told that they would be informed what the total was. The pooled and social loafing replication participants were then shown a box with no dividers, so that their uses fell into a common bin.

Participants who were to be given the standard (individual scores–standard, pooled–standard) next read,

This experiment is a replication of one done last quarter. We would like to see whether or not we can replicate these findings. Since this is a replication, we know the average number of uses generated in this previous experiment and we thought you might be interested in knowing this average. We will give you this information at the end of the session, but please hold it in confidence since it could affect the performance of later participants.

In the individual scores–standard condition, it was made clear that the standard would be the average of individual outputs, whereas in the pooled–standard condition, the participants were told that it would be the average of group performances. In the pooled–no standard condition, participants also read that we knew this average, but were informed that because this information could affect the performances of later participants, we would not be able to share it with them. In the social loafing replication condition, participants read nothing about a standard being provided or withheld.

The experimenter then told participants that they would be generating uses for a knife and reminded them that they were to generate as

many uses as possible without regard to how ordinary or unusual the uses were. They were given 12 min to generate their uses, a time that previous research (Harkins & Petty, 1982) has shown to be ample. At the end of this period, the experimenter came in and gave the participants a questionnaire on which they were to rate various aspects of their experience on 11-point scales. Participants were asked to rate the extent to which the experimenter would be able to tell how well they performed, the extent to which they would be able to tell how well they themselves performed, and the extent to which the other participants would be able to tell exactly how well they performed. In the pooled conditions, participants were also asked to rate to what extent they would be able to tell how well their group performed and to what extent the experimenter would be able to tell how well their group performed. Each of the rating scales that accompanied these questions was anchored by *Not at all* (1) and *Know exactly* (11). While the participants were responding to the questionnaire, the experimenter counted either the number of uses each participant generated (individual scores condition) or the number of uses the group generated (pooled and social loafing replication conditions). After this, feedback was given if called for (individual scores and pooled conditions), and participants were debriefed and dismissed.

Results

We analyzed the data in one-way ANOVAs with four conditions: individual scores–standard, pooled–standard, pooled–no standard, and social loafing replication. We used the group as the unit of analysis and, where appropriate, the Tukey honestly significant difference (Kirk, 1982) as the a posteriori test.

Manipulation checks. Participants in the individual scores condition felt that the experimenter could evaluate them to a greater extent ($M = 8.8$ on an 11-point scale) than did participants in each of the other conditions (pooled–standard $M = 5.0$, pooled–no standard $M = 5.3$, social loafing replication $M = 5.8$, $ps < .05$), $F(3, 60) = 18.3$, $p < .001$. Participants in the individual scores–standard condition also felt that they could tell how well they performed to a greater extent ($M = 7.7$) than did participants in any of the other conditions (pooled–standard $M = 5.1$, pooled–no standard $M = 4.3$, social loafing replication $M = 5.2$, $ps < .05$), $F(3, 60) = 11.4$, $p < .001$. Individual score–standard participants felt that the other participants could tell more about how well they performed ($M = 5.6$) than did participants in the other conditions (pooled–standard $M = 3.3$, pooled–no standard $M = 2.5$, social loafing replication $M = 3.0$, $ps < .05$), $F(3, 60) = 11.1$, $p < .001$. In each of these analyses, the conditions in which outputs were pooled did not differ, $ps > .20$.

Participants in the pooled conditions were asked to rate both to what extent they would be able to evaluate their group's performance and the extent to which the experimenter could do so. Pooled–standard participants reported that they would be able to evaluate their group's performance to a greater extent ($M = 7.4$) than did pooled–no standard or social loafing replication participants (M s = 4.7 and 4.2, respectively, $ps < .05$), $F(2, 45) = 25.4$, $p < .001$. Participants in these conditions felt that the experimenter could evaluate their group's performance to an equivalent extent (pooled–standard $M = 8.3$, pooled–no standard $M = 9.2$, social loafing replication $M = 8.3$, $p > .20$).

Number of uses. Replicating previous research (e.g., Szymanski & Harkins, 1987), participants in the social loafing replication condition generated reliably fewer uses ($M = 19.7$) than

did participants in the individual scores–standard condition ($M = 26.5$, $p < .05$). Learning that there was a standard to which they would not be exposed (pooled–no standard) led to performance equivalent to that in the social loafing replication condition ($M = 20.6$, $p > .20$). However, when told that they would be provided with a standard that would allow group evaluation (pooled–standard), participants generated as many uses ($M = 27.6$) as did individual scores–standard participants ($M = 26.5$, $p > .20$), overall $F(3, 60) = 10.7$, $p < .001$.

Discussion

In both experiments, the opportunity for group evaluation was sufficient to motivate performance to the same extent as the prospect of evaluation by other sources such as the self or the experimenter. In each of these experiments, participants in the pooled conditions reported that they felt that their individual performances could be evaluated by the experimenter, their co-participants, and themselves less than did participants in the individual scores conditions. As important, participants in the pooled conditions did not differ in the extent to which they reported that they could be evaluated by these sources. In each experiment, pooled output participants who were provided with a standard reported that they felt they could evaluate their group's performance to a greater extent than did pooled output participants who were not provided with one.

In Experiment 2, we tested our assumption that the prospect of evaluation by the experimenter, although recognized, would not motivate performance at the group level. In both the pooled–standard and pooled–no standard conditions, there was explicit mention in the instructions of the fact that the experiment was a replication of one done last quarter, and that the experimenter wanted to see if the findings could be replicated. The only difference between these conditions was that in one case the experimenter told the participants that they would be given the average (pooled–standard) and in the other that they would not (pooled–no standard). Because the experimenter had informed all of these participants that at the end of the experiment he would count the total number of uses generated by all of them and tell them what this total was, it should have been apparent to the participants in both the pooled–standard and pooled–no standard conditions that the experimenter could evaluate their group's performance, and the participants in these conditions acknowledged this (pooled–standard $M = 8.3$; pooled–no standard $M = 9.2$, $p > .20$).

We had assumed that in previous loafing research, the participants in the loafing cells had believed that the experimenter could evaluate their group's performance, because it was obvious that he or she would have access to their scores (output) and the scores of the other groups (a social standard). Consistent with this view, the social loafing replication participants indicated that the experimenter could evaluate their group ($M = 8.3$) to the same extent as did participants in the other pooled groups ($ps > .20$). Thus, participants in each of the pooled groups acknowledged that the experimenter could evaluate their group's performance, but nonetheless worked harder only when they could evaluate their group's performance (pooled–standard). Of course, it is quite possible that this outcome would change if the experimenter had more power over the

group's outcomes. An experimenter who could punish the group as a whole might have a very different effect than the experimenter in this research.

Our findings suggest that the prospect of group evaluation is sufficient to motivate performance, whether the evaluation occurs through the comparison of the group's output to an objective standard (optimizing task, Experiment 1) or through its comparison to a social standard (maximizing task, Experiment 2). These findings are consistent with Goethals and Darley's (1987) extension of social comparison theory to the group level. In fact, the setting provided by the loafing paradigm seems ideal for testing Goethals and Darley's notions, because in the loafing cell the possibility of evaluation of individual performances is minimized. By then providing or withholding access to group-level measures of output and a standard, the motivational properties of group evaluation can be examined.

Several other recent formulations have also suggested that the potential for group evaluation could motivate performance. For example, Breckler and Greenwald (1986) have suggested that the self presents itself to three significant evaluative audiences, which give rise to three motivational facets of the self. The public self is sensitive to the evaluation of others. The private self is guided by internal standards that permit self-evaluation in the absence of others. The collective self is the "we" facet of the self and is oriented toward collective achievement—achieving goals and fulfilling one's role in a reference group.

In our research, the operation of the public self can be seen in the finding that the potential for evaluation of individual performance by the experimenter is sufficient to motivate performance (e.g., Szymanski & Harkins, 1987). In the same set of experiments, we found that the potential for self-evaluation was sufficient to motivate performance, consistent with Breckler and Greenwald's (1986) notion of the evaluative potency of the private self. However, we should point out that in this research, the self-evaluation effect was only obtained when an explicit standard was available for comparison, whereas Breckler and Greenwald (1986) alluded to the sufficiency of "internal standards" (p. 148). Certainly, one could have such internal standards, given sufficient experience on a particular task; however, on tasks that are new to the participants, such as ours, this does not seem likely. In any case, in the current research it is possible that the group evaluation effect represents the operation of the third motivational facet, the collective self.

Bandura (1986) has suggested that self-evaluation can occur through one of several types of comparisons: Normative comparisons occur through comparison of one's performance to the performance of a representative group; social comparisons involve the comparison of one's performance to that of particular associates; and self-comparisons entail the comparison of one's output to one's previous behavior. Finally, Bandura (1986) suggested that in societies that are organized around collectivist principles, self-appraisal may be based "on one's relative contribution to the group accomplishment, and how it measures up to the standard adopted by the group" (p. 348). Thus, under some circumstances at least, the opportunity for collective comparison could motivate performance.

In their elaboration of social comparison theory (Festinger, 1954), Levine and Moreland (1986) have developed a typology of outcome comparisons that incorporates three dimensions:

type of comparison, social context, and temporal context. In addition to the self-other comparison that has been the traditional focus of interest in this area, Levine and Moreland added self-self and group-group comparisons. These types of comparisons can take place within the group (intragroup) or across groups (intergroup)—the social contexts proposed by Levine and Moreland—and can occur in the same time period (intra-temporal) or across time periods (intertemporal)—the temporal context dimension of the model. Applying this perspective to the current research, Experiment 2's pooled-standard condition would fall in the group-group, intergroup, intertemporal cell of Levine and Moreland's typology, because participants were afforded the opportunity to compare their group's performance to the average performance of previous, different groups. Experiment 1 does not fit into Levine and Moreland's typology because we used a nonsocial standard. A variety of factors determine the choice of target comparison, which we will return to later, but in any case, the suggestion that group-group comparisons motivate performance is central to their analysis.

Each of these approaches incorporates the notion that the potential for group evaluation can motivate performance, but under what conditions? Breckler and Greenwald (1986) suggested that the collective self's ego task is "achieving the goals of and fulfilling one's role in a reference group" (p. 148), which may include coworkers, religious organizations, clubs, athletic teams, and family. Apparently, then, performing a task in service of such a group would engage this aspect of the self. Bandura (1986) suggested that collective comparison would occur in societies organized around collectivist principles, and that self-appraisal would be based on one's contribution to the group. Thus, these approaches suggest that the opportunity for group evaluation may motivate performance in particular types of groups or cultures.

Because both Levine and Moreland (1986) and Goethals and Darley (1987) worked from social comparison theory, it is not surprising that they have suggested that similar factors lead to a concern about group evaluation. Levine and Moreland (1986, p. 289) suggested that group-group comparisons are likely when

group membership is salient because of group competition, group dissimilarity, or status differentials between groups (Brewer, 1979), when group members seek to enhance their social identity through intergroup comparisons (Tajfel, 1978), or when group members seek to demonstrate their group's inferior outcomes in order to produce social change (Austin, 1977; Patchen, 1961).

Along similar lines, Goethals and Darley (1987) suggested that "comparison is not always sought out, but will take place whenever another group is salient, available, or similar" (p. 33).

An initial examination of these approaches to group evaluation would not necessarily lead one to expect reliable effects in our paradigm. Our "groups" were composed of strangers who only expected to be together for the duration of the experiment. These groups were not the coworkers, religious organizations, clubs, athletic teams, and family to which Breckler and Greenwald (1986) referred. Our society is not one organized around the collectivist principles described by Bandura (1986). The features that Levine and Moreland (1986) and Goethals and Darley (1987) described as making group-group comparison likely also do not seem to be strongly operative in our settings.

For example, there was little in the experimental setting that made group membership salient.

Nonetheless, there were aspects of these settings that could motivate group performance. For example, despite the fact that participants in our groups do not have a past or a future with one another and there was no interaction during the experiment itself, a group product was generated, and, when a standard was provided, there was the potential for group evaluation. Also, it is quite likely that the participants who made up the groups saw themselves as being highly similar to one another. They are about the same age and are likely to see themselves as sharing most values. In addition, although these tasks (e.g., brainstorming, vigilance) have been described as boring, uninteresting, and unlikely to be personally involving (Brickner et al., 1986; Harkins et al., 1980), it is quite possible that the participants saw the tasks as having some minimal value. Beyond these factors, in the present paradigm we minimized the possibility of the evaluation of individual performances by the experimenter, coparticipants, and the self.

Given these minimal cues and the absence of concern for other, perhaps more potent, sources of evaluation, the possibility of group evaluation was sufficient to motivate performance, perhaps for the very reasons suggested by Goethals and Darley (1987). In Experiment 2, in the crucial condition, the groups could compare their output with that of the "average" group from a previous study. In so doing, they could see how their group performed relative to other groups composed of similar participants, thereby providing the opportunity for self-knowledge and self-validation through the group. By outperforming the average group, the participants could not only gain information about themselves through their group's performance, but they could also take pleasure in finding that their group was above average, validating themselves through their group's performance. In Experiment 1, we used a nonsocial, objective standard alone; thus, this experiment represents a test of the motivational potency of the knowledge aspect alone. The group could only compare its performance with an objective standard—the number of signals actually presented—and the results suggest that the promise of this information was sufficient to motivate performance.

These experiments show that even these minimal groups are sufficient to arouse comparison motives, even on these "boring, uninteresting" tasks. In future research, it would be interesting to determine what would be required to reduce the participants' motivational level. For example, if participants were led to believe that their coparticipants were quite dissimilar from them or that the participants whose outputs comprised the social standard were dissimilar from them, one would expect that no group evaluation effect would be observed, because there would be no opportunity for gaining knowledge or validation through the group. In the former case, by making the participants appear quite dissimilar, the participants would no longer feel that they were part of a group; in the latter, the comparison standard would no longer be appropriate. Also, once the group has seen how it can do, motivation may drop because participants may feel that the knowledge and validation functions have been served. Finally, minimizing the importance of the task may also reduce motivation because no useful information could be gained from performing on such a task.

It could be argued that, despite its billing, there is little that is "groupy" about the social loafing paradigm. The pooling manipulation that is central to the effect simply represents one of many alternative means of manipulating evaluation potential. In the paradigm, strangers come together for a brief time with no promise of future interaction. There is little in the procedure itself to make the participants feel that they are part of a group. There is no interaction during the experiment; they are not invited to compete with other groups, or even to try to outdo their own group's last effort. This methodology has suggested potentially interesting approaches to other research problems (e.g., the self-evaluation process; Szymanski & Harkins, 1987), but appears to have contributed little to the group literature.

However, the current research has shown that the loafing paradigm has something to contribute, and interestingly enough, the very features of the paradigm that were cited as weaknesses now appear to be its strengths. That is, it is true that the loafing paradigm incorporates few cues that would lead participants to feel part of a group. Nonetheless, our research indicates that even in this bare bones setting, the promise of group evaluation is sufficient to motivate performance. By focusing on the few remaining cues in this setting, we can learn something about what leads participants to feel that they are part of a group that is worthy of their efforts. Certainly, each of the theories of group evaluation that we have described has much to offer (Bandura, 1986; Breckler & Greenwald, 1986; Goethals & Darley, 1987; Levine & Moreland, 1986), but it appears to take less than these theories would lead one to believe to motivate group performance. The current research begins to suggest how much less.

References

- Austin, W. (1977). Equity theory and social comparison processes. In J. Suls & R. Miller (Eds.), *Social comparison processes: Theoretical and empirical perspectives* (pp. 279-305). Washington, DC: Hemisphere.
- Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice-Hall.
- Breckler, S., & Greenwald, A. (1986). Motivational facets of the self. In R. Sorrentino & T. Higgins (Eds.), *Handbook of motivation and cognition* (pp. 145-164). New York: Guilford Press.
- Brewer, M. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, 86, 307-324.
- Brickner, M., Harkins, S., & Ostrom, T. (1986). Personal involvement: Thought provoking implications for social loafing. *Journal of Personality and Social Psychology*, 51, 763-769.
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117-140.
- Goethals, G., & Darley, J. (1987). Social comparison theory: Self-evaluation and group life. In B. Mullen & G. Goethals (Eds.), *Theories of group behavior* (pp. 21-47). New York: Springer-Verlag.
- Harkins, S. (1987). Social loafing and social facilitation. *Journal of Experimental Social Psychology*, 23, 1-18.
- Harkins, S., & Jackson, J. (1985). The role of evaluation in eliminating social loafing. *Personality and Social Psychology Bulletin*, 11, 457-465.
- Harkins, S., Latané, B., & Williams, K. (1980). Social loafing: Allocating effort or taking it easy? *Journal of Experimental Social Psychology*, 16, 457-465.
- Harkins, S., & Petty, R. (1982). Effects of task difficulty and task uniqueness on social loafing. *Journal of Personality and Social Psychology*, 43, 1214-1229.
- Harkins, S., & Szymanski, K. (1988). Social loafing and self-evaluation with an objective standard. *Journal of Experimental Social Psychology*, 24, 354-365.
- Ingham, A., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology*, 10, 371-384.
- Jackson, J., & Williams, K. (1985). Social loafing on difficult tasks: Working collectively can improve performance. *Journal of Personality and Social Psychology*, 49, 937-942.
- Kerr, N. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology*, 45, 819-828.
- Kerr, N., & Bruun, S. (1981). Ringelmann revisited: Alternative explanations for the social loafing effect. *Personality and Social Psychology Bulletin*, 7, 224-231.
- Kirk, R. (1982). *Experimental design*. Belmont, CA: Wadsworth.
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37, 823-832.
- Levine, J., & Moreland, R. (1986). Outcome comparisons in group contexts: Consequences for the self and others. In R. Schwarzer (Ed.), *Self-related cognitions in anxiety and motivation* (pp. 285-303). Hillsdale, NJ: Erlbaum.
- Patchen, M. (1961). A conceptual framework and some empirical data regarding comparisons of social rewards. *Sociometry*, 24, 136-156.
- Petty, R., Harkins, S., & Williams, K. (1980). The effects of group diffusion of cognitive effort on attitudes: An information-processing view. *Journal of Personality and Social Psychology*, 38, 81-92.
- Steiner, I. (1972). *Group processes and productivity*. New York: Academic Press.
- Szymanski, K., & Harkins, S. (1987). Social loafing and self-evaluation with a social standard. *Journal of Personality and Social Psychology*, 53, 891-897.
- Tajfel, H. (1978). Social categorization, social identity, and social comparison. In H. Tajfel (Ed.), *Differentiation between social groups: Studies in the social psychology of intergroup relations* (pp. 61-76). London: Academic Press.
- Tajfel, H., & Turner, J. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. Austin (Eds.), *Psychology of intergroup relations* (pp. 33-48). Chicago: Nelson-Hall.
- Williams, K., Harkins, S., & Latané, B. (1981). Identifiability as a deterrent to social loafing: Two cheering experiments. *Journal of Personality and Social Psychology*, 40, 303-311.

Received September 9, 1987

Revision received December 12, 1988

Accepted December 13, 1988 ■