Motivation Gains in Performance Groups: Paradigmatic and Theoretical Developments on the Köhler Effect

Article *in* Journal of Personality and Social Psychology · November 2000

DOI: 10.1037/0022-3514.79.4.580 · Source: PubMed

3 authors, including:

Guido Hertel
University of Münster
137 PUBLICATIONS **4,610** CITATIONS

SEE PROFILE

Journal of Personality and Social Psychology
2000, Vol. 79, No. 4, 580–601

0022-3514/00/$5.00 DOI: 10.1037//0022-3514.79.4.580

Motivation Gains in Performance Groups: Paradigmatic and Theoretical Developments on the Köhler Effect

Guido Hertel
University of Kiel

Norbert L. Kerr and Lawrence A. Messé
Michigan State University

In contrast to many demonstrations of social loafing, relatively few studies have documented group motivation gains. One such exception was O. Köhler's (1926, 1927) finding that team members working together did better at a taxing persistence task than would be expected from their individual performances, particularly when there was a moderate discrepancy in coworkers' capabilities. In Experiment 1, we developed a paradigm within which Köhler's overall motivation gain effect could be replicated, although the discrepancy in coworkers' capabilities did not moderate these motivation gains (after statistical artifacts were taken into account). Experiment 2 indicated that this motivation gain occurred under conjunctive but not under additive task demands, suggesting that the instrumentality of one's contribution to valued outcomes is a more likely explanation of the Köhler effect than social comparison processes.

Allport (1924) suggested that the master question of social psychology is how individual behavior is altered in group contexts. And, despite the popularity (hegemony?) of individualistic and cognitive approaches in contemporary social psychology (cf. Moreland, Hogg, & Hains, 1994; Steiner, 1986), most social psychologists would probably agree that—whether or not it is the master question—the comparison of group versus individual processes certainly remains a critical question for our discipline. The present article addresses one aspect of that central question: How does working in a group context alter individual task motivation?

This question not only remains central in present-day social psychology, but, historically, it is the field's first question. Whether one gives the honor to Triplett's (1897) classic study of social facilitation or to the group size studies of Ringelmann (1913, research conducted earlier than Triplett's but published only later; cf. Kravitz & Martin, 1986), the first systematic empirical investigations in scientific social psychology were concerned with the comparison of individual versus group task performance.

Much progress has been made on this general question in the intervening century. The motivational consequences of the mere presence of others is now fairly well understood (Baron, 1986; Zajonc, 1965). Moreover, thanks to Steiner's (1972) seminal work, psychologists now understand that simple comparisons of individual and group performance are rarely very illuminating and that we learn much more of value by comparing actual group performance with the potential productivity of groups, usually estimated with models which assume no motivational differences between individuals and groups and assume optimal use and combination of group member resources. And, over the last quarter of a century (since Ingham, Levinger, Graves, & Peckham's important 1974 article), we have identified a number of group task contexts and psychological processes that lead group members to show what Steiner (1972) called "group motivation losses" and what Latané, Williams, and Harkins (1979) termed "social loafing"—lower task motivation in the group than in the individual performance context (for overviews, see Baron, Kerr, & Miller, 1992; Karau & Williams, 1993; Shepperd, 1993).

In the present article, we present paradigmatic, empirical, and theoretical work on a relatively neglected but important possibility—that there are interesting group contexts within which group members will have higher task motivation than they do as individuals; that is, there are replicable group motivation gain phenomena as well as the more frequently observed group motivation losses.

Group Motivation Gains

Although group motivation gains were demonstrated well over 70 years ago (Köhler, 1926, 1927)—seminal work that was the impetus for the present investigations—contemporary interest in this phenomenon probably grew out of Hackman and Morris's (1975) suggestion that conditions should exist that would foster greater collective than individual effort. However, empirical demonstrations of such motivation gains have been very rare. Perhaps this should not be so surprising. There are good reasons why it should be easier to empirically demonstrate motivation losses than motivation gains. Whenever exerting task effort is costly (e.g., fatiguing, with opportunity costs, etc.), there is an intrinsic incentive to reduce one's effort whenever it is safe or sensible to do so; in contrast, an intrinsic incentive to increase one's effort is less obvious. Moreover, under typical laboratory conditions, individual motivation is already likely to be nearly maximal. Individual

This project was partially supported by a postdoctoral grant from the Deutsche Forschungsgemeinschaft (He 2745/1-1/2).

Correspondence concerning this article should be addressed to Guido Hertel, University of Kiel, Institut für Psychologie, Olshausenstrasse 62, 24 098 Kiel, Germany; or to Norbert L. Kerr or Lawrence A. Messé, Department of Psychology, Psychology Research Building, Michigan State University, East Lansing, MI 48824. Electronic mail may be sent to hertel@psychologie.uni-kiel.de, kerr@pilot.msu.edu, or messe@pilot.msu.edu.

performers are exhorted to "do their best." Usually, their performance is closely monitored and subject to evaluation by the experimenter and potentially by others (Harkins & Szymanski, 1989). And in some instances, good individual performance is also rewarded tangibly. If individuals are already working as hard as they can, we would not expect group contexts to induce gains in task motivation. So, at least one challenge for discovering and analyzing group motivation gains is developing paradigms within which they can occur.

Although clear, well-replicated motivation gains might be the hen's teeth of group performance research, there are a few intriguing findings that might well qualify as sightings:

(a) The facilitation of performance at simple, well-learned tasks in the presence of audiences or coactors (Zajonc, 1965) might be counted as a type of group motivation gain (if it is indeed mediated by enhanced drive; see Baron, 1986, for a competing interpretation).

(b) There are a number of studies that suggest that implicit or explicit competition between members of ostensibly cooperative task groups can enhance member motivation and performance (e.g., Erev, Bornstein, & Galili, 1993; Stroebe, Diehl, & Abakoumkin, 1996, Experiments 2–4). In fact, it is at least as plausible that the "social facilitation" effect reported in Triplett's (1897) classic experimental study was due to implicit competition between children winding fishing reels as to the mere presence of a coactor.

(c) There are a few studies (e.g., Kerr & MacCoun, 1984; Kerr & Sullaway, 1983) that have suggested that group composition can underlie some group motivation gains. In particular, Kerr and his colleagues have found higher task motivation by both male and female participants in mixed-sex groups than in same-sex groups or individual performers, an effect they attributed to special evaluation concerns arising in mixed-sex interactions.

(d) There is some evidence that when difficult performance goals have been set, people may work harder in a group than individually (Matsui, Kakuyama, & Onglatco, 1987).

(e) Williams & Karau's (1991) studies of "social compensation" probably provide the best extant evidence for a genuine group motivation gain. They show that when one's dyad partner is either unwilling or unable to perform well at a task that is very important to one, one will work relatively harder than participants in comparable coacting pairs.

(f) Köhler (1926, 1927) found that people performed a demanding physical persistence task better when working together in dyads or triads than would be expected from their efforts as individuals. This is, as best we can tell, the first published demonstration of a group motivation gain. More important, we suggest that it is a distinctive one and not simply a manifestation of some other motivation gain phenomena (social compensation, social facilitation, etc.). As noted above, successfully replicating and explaining the motivation gain observed by Köhler were the primary objectives of the present studies. Therefore, we examine Köhler's work in some detail below.

The Köhler Effect

Witte (1989) was the first contemporary social psychologist to note Köhler's (1926, 1927) work. Just as sustained research on group motivation losses was stimulated by rediscovery of some long-ignored results—Steiner's (1972) and Ingham et al.'s (1974) discussion of Ringelmann's pioneering studies—it is possible that Witte's review of Köhler's investigations could serve a similar function for the study of group motivation gains. In the studies of most direct relevance to the present research, Köhler asked male rowing club members to perform a simple motor task either as individuals or in dyads. In the individual condition, the rower held a bar connected to a 41-kg weight through a series of pulleys. His task was to do standing bicep curls for as long as possible, paced by a metronome with a 2-s interval. In the dyad condition, the weight was doubled to 82 kg, and one member of the dyad gripped each side of the bar.

Köhler was especially interested in the effects of group ability composition on group performance. His key findings for dyads are reproduced in Figure 1. The x-axis of the figure is the ratio of the individual performances (number of seconds persisting at the task) of the weaker to the stronger dyad member (multiplied by 100 and reversed-scaled for ease of presentation). We refer to this index as the *relative ability ratio* (RAR; see Table 1 for frequently used acronyms and definitions). When RAR = 100, the dyad members did equally well when they performed individually; when RAR = 67, the weaker dyad member lasted only two thirds as long as the stronger member during individual trials; and so forth. The y-axis of the figure is the ratio of the dyad's performance (again, number of seconds persisting at the task) to the average of the dyad members' individual performances (again multiplied by 100 for ease of presentation). We refer to this ratio of Köhler's as the *additive baseline ratio* (ABR) because it effectively assumes that the lifting task is what Steiner (1972), called an additive—or, to be more precise here, an averaging—task. If ABR = 100, it means that the dyad does exactly as well as the average of its members' individual scores; when ABR > 100, the dyad does better than the average of individual scores; when ABR < 100, the dyad does worse than the average of individual scores.

Köhler (1926) suggested that the function relating RAR and ABR in his data set was a curvilinear one. When there was either very little discrepancy in the abilities of the dyad members (RAR > 80) or a very large discrepancy (RAR < 50), the dyads did worse than the average member, whereas for moderate levels of discrepancy (50 < RAR < 80), the dyads did better than the average member. Köhler (and Witte, 1989) took the latter result as evidence for a group motivation gain. (Köhler reported qualitatively similar results for performance triads working at the lifting task and for dyads at a wheel-turning task, although he found no such effects for a group rope-pulling task.) Polynomial regression reanalysis of Köhler's plotted data indeed confirms what he determined through visual inspection, that the function relating RAR (centered to prevent multicollinearity; cf. Neter, Kutner, Nachtsheim, & Wasserman, 1996) and ABR in his data is not a monotonic one. The quadratic component was highly significant, $\beta = -0.797$, $p < .001$ (linear component: $\beta = -0.175$, $p < .16$); $R^2 = 0.68$ (adjusted $R^2 = 0.66$).

Actually, as Stroebe et al. (1996) noted, for purposes of detecting motivation gain, per se, Köhler used an inappropriate baseline of dyad potential productivity, although the ABR might be quite appropriate for other questions (e.g., the design question of whether pairs of individuals are, overall, more or less productive than a yoked-dyad). Köhler's lifting task was really what Steiner (1972) termed a *conjunctive* task, one at which the group can do no

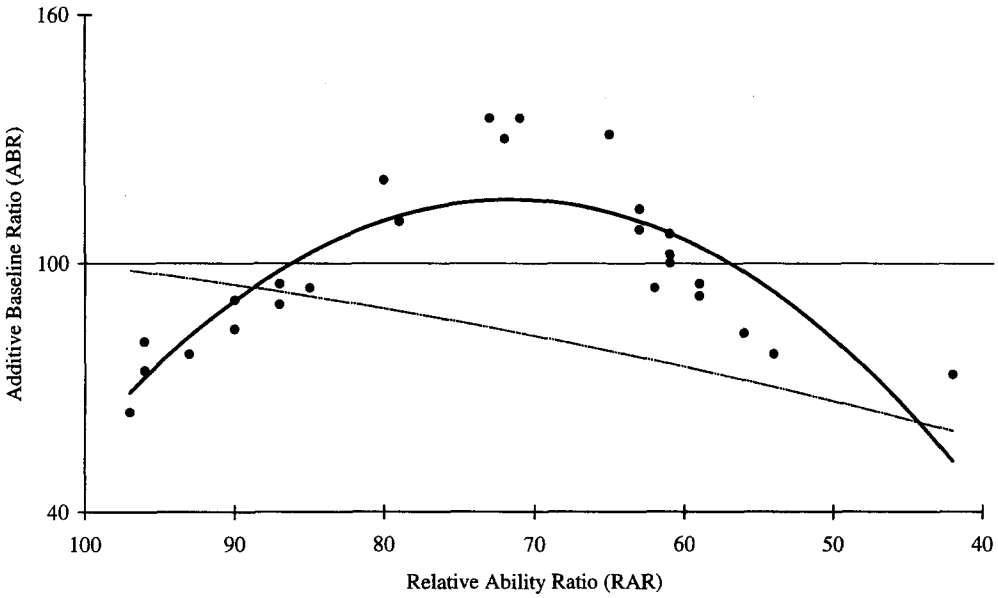

Figure 1. Reproduction of Köhler's (1926) performance results in his dyadic weight lifting tasks. The solid curve represents the quadratic trend of the data based on our reanalysis of Köhler's original results. The horizontal solid line (that crosses the *y*-axis at 100) represents the potential productivity baseline based on the average member's score. The lower, dashed curve represents the potential productivity baseline using the less capable rather than the average member's score.

better than its least capable member. Remember that in Köhler's studies, twice as much weight is being lifted by the dyad as by an individual. As soon as either dyad member stops, the task for the remaining (partially spent) partner immediately becomes twice as difficult and effectively impossible for the stronger partner to continue performing. Thus, if dyad members simply worked exactly as hard in the dyad as they did individually (no motivation loss or gain), the dyad should perform only as well as the weaker member, not as well as the average of the two individual dyad members.

Table 1
Table of Acronyms

Acronym	Conceptual definition
RAR	The ratio of the individual performances of the weaker to the stronger dyad member ([lower individual dyad score/higher individual dyad score] × 100). An index of the discrepancy of dyad member capability.
ABR	The ratio of the dyad's performance to the average of the dyad members' individual performances ([(dyad score)/(average of individual scores of dyad members)] × 100). A faulty index of group motivation gain for a conjunctive task (although useful for gauging group vs. individual efficiency).
CDS	The signed difference between the dyad performance and the worse of the individual dyad members' performances ([dyad performance] − [MIN(individual scores of dyad members]). An index of group motivation gain for a conjunctive task.
ADS	The absolute value of the difference between the dyad members' individual performances (|[MAX(individual scores of dyad members)] − [MIN(individual scores of dyad members)]|). An index of the discrepancy of dyad member capability.

Note. RAR = relative ability ratio; ABR = additive baseline ratio; CDS = conjunctive difference score; ADS = absolute difference score; MAX = maximum; MIN = minimum.

It is not difficult to show that in Köhler's plots, using the less capable rather than the average member's score would mean comparing Köhler's observed curve not with the horizontal line ABR = 100, but rather the curve ABR = 100 × 2 (RAR/RAR + 1).[1] This function is superimposed (the dashed curve) on Köhler's data in Figure 1. Comparing Köhler's index of group performance, the ABR, with the more appropriate potential productivity baseline leads to a slight qualification of Köhler's conclusions. Except for dyads whose members were nearly identical in ability, Köhler's dyads generally outperformed their weaker member. Thus, the evidence for motivation gains in Köhler's data may have been stronger and more general than he realized. Moreover, because Köhler's task also seems very vulnerable to group coordination losses (i.e., reduced group performance due to imperfect coordination of member lifting), if anything, Figure 1 probably underestimates the magnitude of the real motivation gain. As Köhler concluded, however, in his data this motivation gain does appear to be maximal when the discrepancy of member abilities is moderate, RAR ≈ 70. In summary, Köhler's original "effect" was really two effects: (a) an overall motivation gain for his task and (b) the

[1] Let a and b be the scores of the individual dyad members. Further, suppose that $a < b$; that is, Dyad Member A is the less capable of the pair. If the group did exactly as well as its less capable member, A, then the dyad's performance should also be a, and the expected ABR scores, ABR', should just be $a/([a + b]/2)$. Dividing numerator and denominator by b and recalling that RAR = a/b, one concludes that $ABR' = 2(\text{RAR}/[\text{RAR} + 1])$.

moderation of this motivation gain by the discrepancy of member abilities (with maximum gain occurring when discrepancies were moderate). In the present article, we are primarily concerned with the first of these effects.

Replication of the Köhler Effect

To our knowledge, the only extant attempts to replicate Köhler's (1926, 1927) findings are the five experiments reported in Stroebe et al. (1996). The first study attempted to replicate the effect using Köhler's original lifting task. It was successful; that is, dyads did better than their average (Köhler's original baseline) and their less capable member (the appropriate baseline for detecting motivation gains) when there was a relatively large discrepancy in abilities. However, this study also confirmed a serious concern about the lifting task—it is altogether too taxing and too hazardous. Stroebe et al. (1996) reported, "Most of our subjects suffered from intense muscle pain after the first (individual) session and were rather unwilling to participate in the second (group) phase of the experiment" (p. 52, parenthetical comments added). Although cash inducements prompted enough participants to return to enable dyad versus individual comparisons, Stroebe et al. recognized that Köhler's experimental task, acceptable to athletes in the 1920s, is probably not acceptable to today's student participants (or committees charged with protecting the welfare of human participants) and, as such, a different laboratory task would be required to study the Köhler effect.

Stroebe et al. (1996) used one such alternative task in their next three experiments. In a variant of Köhler's second task, participants turned a crank (with a mechanical brake) as fast as possible for 10 min. On all trials, participants worked in separate rooms. To capture the conjunctive aspect of Köhler's task, participants were told that unless the turning speeds of the two dyad members were sufficiently close to one another, a penalty would be assessed. A computer screen continuously displayed the discrepancy in turning speeds between dyad members on dyadic trials.

There was some evidence of motivation gains at this task—dyads generally did better than isolated individuals, and in one study (Experiment 2) motivation gains relative to the proper, weaker-member baseline were positively related to the discrepancy between group member's individual performances. However, dyadic performance was not consistently related to intermember discrepancy. Stroebe et al. concluded that the Köhler effect could not be examined directly with their crank-turning task because certain of its features (viz. the requirement of maximizing cranking speed while having continuous feedback of one another's performance) may have engaged a different motivational process—intermember competition—that masked the Köhler effect.

Finally, Stroebe et al. (1996) described an unpublished thesis by Ruess (1992). In this study, the participant's task was to sit in a chair, attach a 1-kg weight to his or her arm, and then hold the arm horizontally for as long as possible. The arm was held above a string connected at each end to stands at a height of 1 m above the floor. A trial ended when the arm was lowered and broke the string. In the dyad condition, 2 participants held their arms above a single string. Participants participated in two sessions, one assessing individual performance and a second assessing performance in dyads, with order of the sessions alternated.

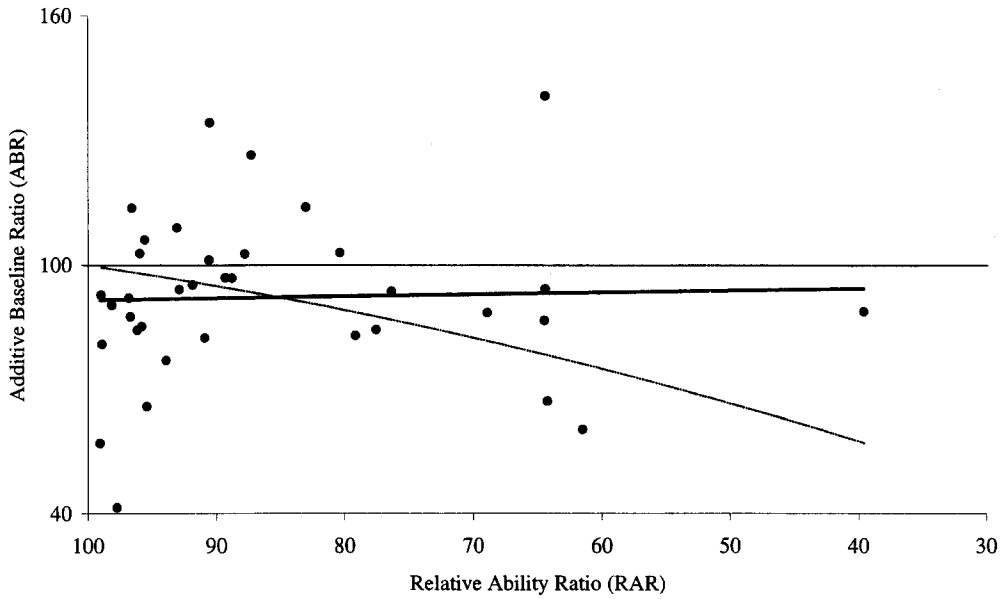

Figure 2. Replotting of Ruess's (1992) performance data using the additive baseline ratio. The solid curve represents the linear trend of the data based on our reanalysis of Ruess's original results. The horizontal solid line (that crosses the *y*-axis at 100) represents the potential productivity baseline based on the average member's score. The lower, dashed curve represents the potential productivity baseline using the less capable rather than the average member's score.

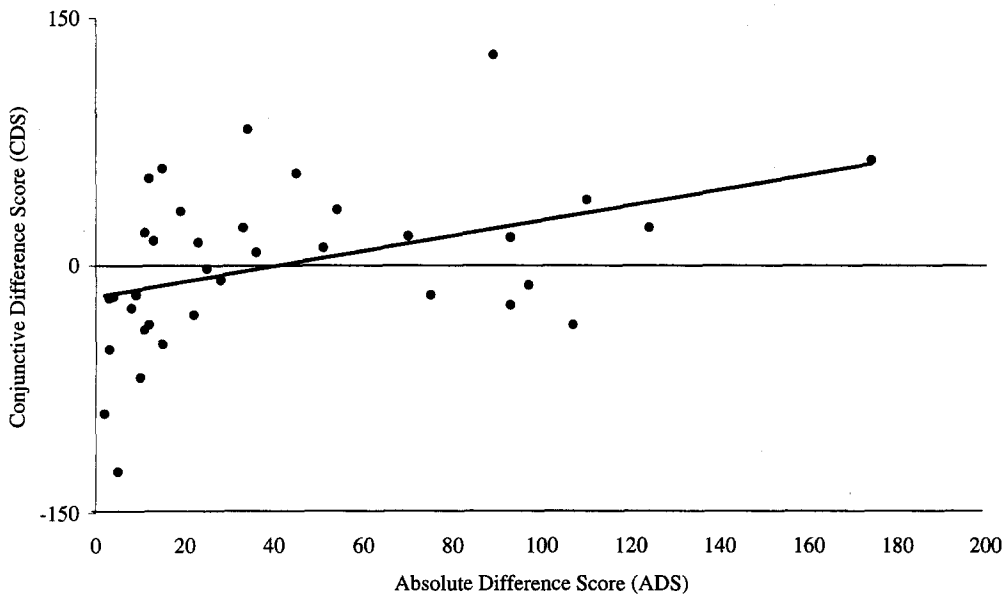

Figure 3. Replotting of Ruess's (1992) performance data using difference scores.

Figure 2 plots Ruess's (1992) data using Köhler's method of presentation.[2] As in Köhler's study, as the discrepancy between dyad members increased, apparently so did the performance of the groups. A polynomial regression analysis of the ABR measure, though, suggested no significant departure from a simple linear function (plotted in the figure). Comparison with the proper, conjunctive baseline (the dashed curve in Figure 2) suggests that Ruess, like Köhler, may have obtained evidence for a genuine motivation gain for dyads whose members' abilities differed sufficiently.

However, plotting the data in this way can be misleading. This potential interpretive problem can be seen more easily in Figure 3, an alternative plot of Ruess's (1992) data. In this graph, the *y*-axis is the difference between the dyad performance and the worse individual performance, a variable we call the *conjunctive difference score* (CDS). This index has at least three advantages over the ABR that Köhler used. First, it uses the proper no-motivation-gain baseline, that of the less capable individual, rather than the average of the dyad members' individual scores. Second, it is more informative: The magnitude of the CDS expresses just how much better or worse the dyad did relative to its less capable member in the task's unit of performance (e.g., for Ruess, 1992, that unit is the number of seconds participants could hold their arm above the string). This advantage also suggests that it would be useful as well to index the discrepancy in dyad members' scores with an *absolute difference score* (ADS) in Figure 3, rather than the ratio of more to less capable member score (the RAR). Third, the CDS is not biased by regression-to-the-mean artifacts, whereas the ABR is.[3]

Reanalysis of Ruess's (1992) data revealed two noteworthy results. First, there was no net motivation gain across all participants. The overall mean CDS was 0.78 s, which did not differ significantly from the zero baseline, $t(35) = 0.1$, *ns*. Second, Köhler's simple inverted-U function relating discrepancy in dyad members' performances to relative dyadic performance was not replicated.[4] Polynomial regression analyses on Ruess's (centered) data showed that a quadratic solution did not substantially improve the amount of explained variance (linear component: $\beta = 0.599$, $p < .02$; quadratic component: $\beta = -0.264$, $p > .25$; $R^2 = 0.20$; adjusted $R^2 = 0.15$) over a linear solution ($R^2 = 0.16$; adjusted $R^2 = 0.14$). Clearly, the data are not well fit by a simple inverted-U function. Figure 3 shows the best fitting linear function (the solid line; $\beta = 0.405$, $p < .02$). The basic pattern in the data was that dyads improved (relative to their less capable member) as member discrepancy increased.

However, it turns out that the same regression artifacts that bias the ABR index of dyad performance also bias the relationship between member discrepancy (ADS) and dyad performance (CDS). We postpone a fuller discussion of this bias for the mo-

[2] We wish to express our sincere thanks to M. Ruess for providing us with his data.

[3] This attribute can be demonstrated via simulation or logically. We present simulation results a bit later in this article. Logically, both Köhler's (1926, 1927) and Ruess's (1992) tasks require that the dyad's score be defined by which member quits first, that is, which member gets a more extremely low score when tested in the dyad. As long as there is error of measurement, extreme scores (here, extremely low scores) will regress toward the mean in independent testings. Thus, whenever poor dyad performance is the consequence of error, we would expect that member's score to be relatively higher in an independent testing (e.g., when being tested as an individual) and for the ABR's denominator to also be relatively higher, and thus for the resultant ABR ratio to become smaller. So even if the dyad does perform at the level of the less capable member, on retesting as an individual, that member will, on average, tend to have a higher score, which will, overall, result in a mean CBR < 1.0.

[4] We note that Stroebe et al. (1996) reported an analysis that concludes that Ruess (1992) did replicate the inverted-U. However, those analyses are somewhat equivocal. For example, they did not weight all data points equally; observations at large discrepancies were functionally given larger weights than observations with small discrepancies.

ment. It is sufficient for now to note that the magnitude of this relationship in Ruess's data (linear $r = .405$) is well within the range of what one would expect by chance alone (see Appendix A). Thus, Ruess provided little clear evidence of either a motivation gain or of its moderation by the discrepancy of member abilities.

Summary and Overview of Objectives

Köhler's (1926, 1927) early experimental work may have documented a genuine group motivation gain effect. The effect appears to be robust when using Köhler's original task and paradigm (Stroebe et al., 1996, Experiment 1), but both that task and his paradigm more generally have serious drawbacks. Besides being inefficient, they may pose unacceptable risks of pain and/or injury to participants. Alternative paradigms (Stroebe et al., 1996, Experiments 2–5) have failed to replicate Köhler's findings. Thus, the next step in exploring the Köhler effect should be to develop an efficient and safe experimental paradigm that both avoids these drawbacks and replicates the motivation gain effect. This was the primary objective of Experiment 1. Then, in Experiment 2, we again used that paradigm to competitively test two alternative explanations for Köhler's motivation gain.

Experiment 1

Ideally, the new experimental paradigm we sought should incorporate as many features of Köhler's original methodology as possible. This would minimize the chances of omitting any feature, as yet unrecognized, that could be essential to the psychological processes underlying Köhler's results. There were two features of the task, however, that we wanted to modify: (a) it must not pose any risk of physical injury or distress for participants, and (b) it should require little, if any, coordination between dyad members, so as to minimize the likelihood that coordination losses would obscure any motivation gain effects (cf. Ingham et al., 1974; Steiner, 1972). Moreover, for Köhler's task, there is at least a possibility that the nature of the yoking simplified the task for the less capable member—a kind of "coordination gain." This possibility seems a bit remote, but would be effectively eliminated by minimizing the task's coordination requirements.)

Potentially Essential Features of a Task Paradigm to Study the Köhler Effect

Certain desirable features are immediately evident. First, (1) the task should be one for which performance is monotonically related to effort; that is, it should never be possible to "try too hard," such that an increase in effort could lead to lower performance. In that same vein, it would be preferable that (2) performance is primarily dependent on level of effort, per se, and not on such factors as training, skill, intelligence, or related aptitudes. These requirements recommend a simple physical task, like the lifting task used by Köhler. Next, (3) the task ideally should have no functional ceiling on performance, or, if such a feature is impossible to achieve, that ceiling should be as "high" as possible. By this we mean that even when individual performers honestly feel they are "doing their best," there should still be room for improvement of performance and, hence, at least a possibility of a motivation gain. For our present purposes, what is desirable is an experimental situation in which the average individual participant's "best" performance is actually below his or her true optimal performance (i.e., under ideal conditions). This gap will leave room for the improvement required to detect a genuine group motivation gain effect.

Then, because we wanted to explore the relationship between ability discrepancies and group-member motivation, (4) it is essential that there be some variability in individual capacities—and the greater the variability the better. Also, as a matter of efficiency, (5) it would be desirable if several task trials, preferably in different dyads with different relative abilities, could be completed in a single experimental session (without exhausting or injuring participants, of course).

There are a number of additional features inherent in Köhler's lifting task that we sought to duplicate. Specifically, in Köhler's task: (6) Dyad members were physically "yoked" at a conjunctive task. When one stopped, the other could not realistically continue. (7) Dyad members were in one another's physical presence. (8) The task was a taxing persistence task; the performance criterion was how long one could persist. (9) It was probably evident to either dyad member when the other was having difficulty persisting. On the other hand, after coming to this realization, (10) there was probably still time for group members to react. (Together, these last two features suggest that persistence at Köhler's lifting task was not an immediate, "all or none" matter, but one in which there were growing indications of who was and wasn't nearing his performance limit and time for members to act on that knowledge.) (11) Both individuals and groups cared about performing well and about demonstrating their competence to one another. And, although it is not completely clear from Köhler's description of his method, it also seems probable that (12) individual and dyadic performance was observable not only to the experimenter, but to other members of a superordinate group (the rowing club), (13) a group with which participants strongly identified.

They may be self-evident, but at this point it might be useful to explicitly note two important points. First, it is possible that there are additional features of Köhler's paradigm that are either necessary or highly facilitative for the motivation gain effect he observed. For example, it appears that Köhler's group members were able to freely communicate with one another during performance trials, a feature that we decided to omit for reasons of experimental control (see below). If we have overlooked such a feature, it ought to be difficult or impossible to replicate his results in our own paradigm, thereby necessitating further modification and testing. Second, we are not arguing that all 13 of the features that we have identified above are necessary to produce the motivation gain that Köhler observed, and, thus, this effect only occurs in those (possibly few) situations where all (or even most) of these elements are present. We have included certain features (e.g., #1, #2) for purely methodological reasons, and it could turn out that relatively few of the remaining features we have identified are essential to the phenomenon. We are arguing, though, that one effective way to distinguish the essential from the inessential aspects of Köhler's context is to begin with a performance setting that contains as many potentially essential features as possible. If we could achieve this in Experiment 1 and successfully replicate Köhler's motivation gain, we could then proceed (in Experiment 2 and any subsequent studies) to systematically vary the presence or levels of

those contextual features that contending theories identify as crucial.

A Paradigm

Ruess's (1992) promising arm-above-the-string paradigm developed and described by Stroebe et al. (1996) was our starting point for creating a useful parallel method to Köhler's. Ruess's task appears to incorporate many of the essential features noted above. In particular, it is a simple physical persistence task for which effort and performance should be closely related (#1, #2, & #8 from the previous listing); participants can work on the task in one another's presence and, thus, can observe one another's performance (#7 & #12); it can generate substantial between-individual variability in ability (Ruess's dyads had RARs [weaker–stronger individual scores] ranging from 100 to less than 65; #4); and, because the trial ends whenever either dyad member's arm hits the string, the task demands are conjunctive (#6). Most important, although it may pose risks of mild fatigue and muscular soreness, the Ruess task does not generate the substantial risks of serious injury or exhaustion implicit in the Köhler lifting task. Further, there is practically no coordination required of dyad members when performing this task, minimizing the probability of coordination losses.

Other desired features could be attained through simple modifications of Ruess's (1992) task. The physical yoking of dyad members (#6) can be achieved by having dyads grasp a single weighted bar instead of each member's having a separate weight attached to his or her wrist. And, also consistent with Köhler's task, individuals could work with a bar weighing half as much as the bar dyads hold. It appears (Diehl, personal communication, August 3, 1997) that Ruess's participants held their arms a relatively short distance over the string. This arrangement reduces the chance that group members (e.g., the less capable member) could tell when one member was approaching his limit in time to do anything with this knowledge (e.g., decide not to quit just yet). To better approximate Köhler's conditions, we decided to have participants (both individuals and dyads) hold their bar 10 inches (or about 25 cm) above a trip rod. With this arrangement, as soon as either dyad member started getting fatigued and his or her arm started to drop, it would be immediately evident to both members (through the slope of the bar; #9), and both could become aware of this eventuality well before the trial ended (i.e., before the weaker member's arm dropped all the way to the trip rod; #10).

In his experiment, Ruess (1992) asked his participants to perform only a single trial with each arm (one individually, one in a dyad). To achieve multiple, nonexhaustive trials per session (#5), we shortened the period participants were likely to persist, partially by increasing the weight (from 1 kg used by Ruess's male participants to 1.55-kg bars for our individual male participants and partially by adding the requirement of grasping the bar and keeping it level. Pilot work indicated that these modifications resulted in shorter performance trials (about 2–3 min/trial vs. about 4–5 min/trial in Ruess's study), and, with sufficient intertrial rest periods, it was feasible to conduct three trials per participant with each arm during a 1-hr session without risking injury or exhaustion.

This increase in task difficulty and in number of trials increased the risk that our individual participants might be operating at or near their true performance limits, and hence, might not leave dyads sufficient room to show motivation gains (#3). An instructional device was used to reduce this risk. Consistent with the requirements imposed by our Institutional Review Board (human participants committee), participants were told that although they were always to do their best, they were not expected to persist so long as to risk injury or undue fatigue. They were instructed that each had to decide individually how long to persist before the resultant muscular soreness and fatigue became uncomfortable enough to end a trial. This instruction, in addition to safeguarding the health and welfare of the participants, sets the operative performance ceiling at the individual's level of initial discomfort, which is likely to be well below her or his true performance ceiling. This procedure reflects our belief that "doing one's best" at this or other taxing persistence tasks is functionally a matter of deciding just how much discomfort one is willing to accept.

Finally, we wanted to create a task situation in which participants cared about both performing well (#11) and about their groups (#13). Initial pilot work with the modified Ruess (1992) task suggested that task motivation of student participants was not particularly high, even with cash incentives for good performance and experimenter observation and evaluation of performance. To increase task motivation and group identification, we developed the following cover story. The experiment was described as a study of the relative persistence of men versus women. It was noted that such comparisons are only meaningful if the task is of equal difficulty for men and women. Then, it was explained that the weights of the bars used by men and women in our experiment had been chosen to equalize the mean performance scores for men and women. (This statement was quite true. Through pilot work, we determined that to match the difficulty that a 1.55-kg bar posed for the average male participant, the bar for individual female participants should be .79 kg. The bars for male and female dyads were exactly twice as long and twice as heavy as the bars individual men and women used. This arrangement permitted examining both men and women in our modified paradigm.) There was an explicit cash incentive for maximizing individual and group task performance (see below), but, in addition, it was explained that a special bonus would be paid if, on average, one's sex did better overall than the opposite sex. Although this "battle-of-the-sexes" procedure was unlikely to completely reproduce the levels of task importance and group identification of Köhler's rowers, it was designed to better approximate these features (#11 & #13).

Summary and Objectives

Ruess's (1992) task and procedures were adapted to provide an experimental paradigm that better approximated Köhler's original paradigm. The primary objective of Experiment 1 was to see if Köhler's interesting, but elusive (Stroebe et al., 1996, Experiments 2–5), motivation gain effect could be replicated in this new, hybrid paradigm. Besides our primary focus on task motivation, we also collected a number of subjective reactions of participants to their performance as individuals and in dyads. The impetus for including these additional measures was Köhler's informal observation that his motivation gains were not accompanied by subjective reports of higher fatigue. In fact, Köhler (1926, 1927) reported that members of his moderate-discrepancy groups—the groups that

most exceeded their potential productivity—actually described themselves as being less fatigued than the low-discrepancy groups.

Method

Participants

The participants in Experiment 1 were 84 undergraduates (48 women, 36 men) at Michigan State University who participated for extra course credit. Data were collected in groups of 6 persons. In each session, gender was the same for all participants and the two experimenters, thereby avoiding any possible effects on motivation of participating in mixed-sex contexts (e.g., Kerr & MacCoun, 1984).

Experimental Task and Measures

As noted above, we used a modified version of the persistence task developed by Ruess (1992; see also Stroebe et al., 1996, Experiment 5). Participants were instructed to hold a metal bar in one hand approximately 25 cm above a flexiglas rod for as long as they felt comfortable with the task. This "trip rod" was 130 cm long and was placed on top of two electric switches that were mounted on two wooden poles (each 97 cm high). The electric switches were connected to two red flashlights that lit up as soon as a participant's arm touched the flexiglas bar. Because very little if any coordination of effort was required for this task, we assumed that performance (i.e., persistence at the metal bar-holding task) directly assessed level of effort or motivation. More specifically, performance was defined as the total number of seconds between the start of a trial and when an arm hit the flexiglas bar, measured with a stopwatch by an experimenter.

Individual and dyadic trials were conducted in different small rooms (approximately 2.6 × 2.0 m) off a larger central room, so that a dyad and an individual could be tested simultaneously. To prevent any sounds from one room being heard, and thus possibly affecting participants in the other room, the doors of the rooms were closed during the trials, and polystyrene cushions were placed on the ground below the flexiglas bar to muffle any noise that a trip rod or bar may have made hitting the floor.

During the performance trials, there were two stools in the dyad room and only one in the individual room. The bar in the dyad room was twice as long and as heavy as the bar in the individual room. For men, the bars were 92 and 184 cm long, 3.3 cm in diameter, and weighed 1.55 and 3.1 kg; for women, the lengths were 72 and 144 cm, the diameter 2.9 cm, and the weights were .79 and 1.58 kg. In all other regards, the equipment in the individual and dyad rooms was identical.

In both rooms, participants were asked to place their dominant (or nondominant) hand at marked spots on the bar. In the dyad condition, these marks were 92 cm apart, each 46 cm from the middle of the bar. In the individual condition, the mark was at the center of the bar.

During each trial, both the experimenter and one of the participants silently watched the (individual or dyad) perform the task. This feature was included for two reasons. First, social facilitation effects were controlled, because all performers had two passive observers (the experimenter and another participant). Second, Köhler's (1926, 1927) performers could also be observed by other group (i.e., rowing club) members as they performed the task. It was unclear from Köhler's description of methods whether observers could communicate with and encourage performers during their trials. However, he did mention that some of the participants did not react to the cheering of their coworker (Köhler, 1926, p. 279), suggesting that at least within the teams, such encouragement was allowed. Our own pilot testing found that only a very few observers who were explicitly permitted to communicate actually did so, and those rare communications between dyad partners, when they did occur, mostly involved agreements to quit. Because of the rarity of these communications, our inability to control their content, and indications that they might be used for purposes other than mutual encouragement, we decided to prohibit any verbal communications among participants (i.e., the performer[s] and the observer) during the task.

After each trial, participants also completed a short questionnaire comprised of four 9-point scales: "How strenuous was the last trial?" (1 = *not strenuous*, 9 = *very strenuous*), "How much effort did you put into this last trial?" (1 = *not very much*, 9 = *very much*), "How much fun did you have in this last trial?" (1 = *not very much*, 9 = *very much*), and "How important was it to you to perform well in this last trial?" (1 = *not at all*, 9 = *extremely*).

Procedure

Prior to each experimental session, an experimenter made sure that none of the participants had any disabling arm, shoulder, or back injuries; if necessary, any participant so identified was exchanged with one scheduled for another study that was being conducted at the same time. Participants were also asked to put their watches in their pockets or purses until the end of the session. The experimenter explained in each session that the participants made up a six-woman (or six-man) team; theirs would be called the "Blue Team" in order to distinguish it from other teams. In addition, each of the group members received a blue name tag with the identifying letters A–F. Participants wrote their names on their tags and fixed them on their shirts. All of these steps were taken to maximize feelings of group identification (e.g., Gaertner, Mann, Murrel, & Dovidio, 1989).

Participants were then told that the focus of the study was to compare the persistence of men and women at physical tasks. Therefore, the average performance of six-person male teams would be compared with the average performance of six-person female teams. Participants were informed that different weights were used for male and female groups to adjust for differences in mean body strength, to ensure that the task was equally difficult for the average man and woman. Then the experimental task was explained in more detail. Participants would perform the persistence task both individually and in dyads. It was stressed that in the dyad condition the trial would be over as soon as either one of the dyad members touched the flexiglas bar. Each participant would perform several trials, but the exact number was not revealed.

To further increase the meaningfulness of the task, it was explained that the six-person team could also earn money through its performance on the persistence task. One such team would be randomly picked at the end of the experiment and would receive five cents for every second of the total group performance (TGP) score (up to a maximum of $240). This TGP score was simply the total of all individual and dyad performance times during the session.[5] The total amount of money earned would be split equally among the six team members. In addition, it was explained that a 15% bonus would be paid to the winning team if their own gender did better on average at the persistence task than did the other gender. Thus, high performance at the task not only increased one's payoff if one's group was randomly chosen, but it also could indirectly increase the payoff of some other randomly chosen group of the same gender. With this bonus incentive, the winning team could earn up to $276. It was emphasized that participants should do their best but that they also should not continue if their arms ever became too tired or if the task was causing them too much discomfort.

Before the trial sequence began, the experimenters supervised participants as they adjusted the heights of their stools (adjustable to any height from 45 to 72 cm) so that all participants' horizontally outstretched arms were at the same height (109 cm from the floor). Then the trials were started according to a fixed schedule (see Table 2) that was the same for

[5] Note that due to this algorithm, each performance second in an individual trial was worth twice as much to the six-person group as in a dyad trial, providing an even more conservative test of motivation gains in groups.

Table 2
Trial Schedule of Experiment 1

	Dyad trials (Room 2)		Individual trials (Room 1)		
Trial–arm	Performing	Audience	Performing	Audience	Resting
1–Dom.	A, B	D	C	F	E
2–Dom.	D, E	A	F	B	C
3–Nondom.	A, C	F	B	E	D
4–Nondom.	D, F	B	E	C	A
5–Dom.	B, C	E	A	D	F
6–Dom.	E, F	A	D	C	B
7–Nondom.	A, B	E	C	D	F
8–Nondom.	D, E	C	F	A	B
9–Dom.	A, C	D	B	F	E
10–Dom.	D, F	C	E	B	A
11–Nondom.	B, C	F	A	E	D
12–Nondom.	E, F	B	D	A	C

Note. Letters A–F stand for the 6 participants in each session. Dom. = dominant arm; Nondom. = nondominant arm.

each session. One dyad and one individual trial were conducted at the same time in the two smaller rooms with one experimenter and one silent "observer" in each room. Thus, 3 participants performed at the same time while 2 observed; the remaining participant rested in the central room. This schedule guaranteed that every participant had a rest period between trials on which they used the same arm. Each participant completed one individual and two dyad trials with each arm. Thus, there were three unique sequences for this series of dyadic (D) and individual (I) performance trials: DDI, DID, and IDD. In a session, 2 participants followed each of these sequences for each arm (dominant and nondominant). This schedule resulted in a total of 12 dyad and 12 individual trials in each experimental session (see Table 2). In addition, the members of the two-person male and female teams of research assistants counterbalanced the experimenter roles (i.e., testing the dyads vs. testing the individuals) so that there was no confounding of experimenter with condition (individual vs. dyad trial).

At the beginning of a trial, each worker participant placed his or her stool so that the wrist of the dominant or nondominant hand was (with arm straight and elbow locked) just above the flexiglas bar. The experimenter lifted the metal bar into position approximately 25 cm above and parallel to the flexiglas bar; this position was indicated by wooden dowels attached to and extending above the support stands. After the participants grasped the bar, the experimenter would say "ready, set, go," release the bar, and then start a stopwatch. Participants were instructed to keep their feet on the floor while performing the task. Verbal communication was not allowed during the trial. After each trial, participants completed the short questionnaire measuring their subjective experiences during the trial.

After finishing a session, participants were debriefed and thanked. At the end of the experiment, one group was randomly picked to receive a monetary payoff, as had been promised in the experimental instructions.

Results

Performance Measure

The primary dependent variable was the CDS, the difference between the dyad performance time and the shorter of the times of the dyad members on their individual trials. It was chosen for its several advantages over Köhler's ABR, as discussed earlier.[6]

Preliminary Analyses of Individual Performance Data

As a preface to looking for the Köhler effect, we first analyzed individual trial performance. This was done primarily to determine whether there were fatigue effects; that is, might we expect performance on later trials to decrease simply as a result of fatigue? In our paradigm, fatigue would be evident to the extent that performance at the individual trial depended on how many previous (dyad) trials (none, one, or two) one had completed with the same arm. If there was evidence for such a fatigue effect, estimates of its size could provide a means of correcting for fatigue in the later motivation-gain analyses. In addition, we wanted to determine if there were any reliable performance effects attributable to participant sex and arm used.

We first analyzed individual performance times for effects of sex of the participants, arm (dominant vs. nondominant), and ordinal position during the session (i.e., whether the individual trial was the participant's first, second, or third performance trial using that arm). The $2 \times 3 \times 2$ analysis of variance (ANOVA) resulted only in a significant main effect for the position factor, $F(2, 156) = 6.12, p < .005$, indicating a clear fatigue effect. When the individual trial was the participant's first performance trial for an arm, performance times were higher ($M = 168$ s) than when it was the second ($M = 151$ s) or third ($M = 134$ s) trial. Participants did perform slightly better with their dominant arm ($M = 157$ s) than with their nondominant arm ($M = 145$ s), but this difference was not statistically significant, $F(1, 156) = 2.42, p < .13$. In addition, there was no effect of participants' sex, $F < 1$, revealing that our selection of the bar weights for men versus women was appropriate. More important, there were also no significant interaction effects, all Fs < 1, suggesting that the observed fatigue effects were comparable for men and women, as well as for dominant and nondominant arms. Note that in the preceding analysis, individual performance scores with the dominant and nondominant arm were treated as independent, even though they were correlated, $r = .635, n = 84, p < .001$. Examination of Table 1 reveals that ordinal position was always different for the two arms, and, thus, these two factors could not be crossed in the preceding analysis. However, separate analyses for dominant and nondominant arms, which avoid this assumption of independence, produced essentially the same fatigue effects.

In order to test whether the group trials had some kind of "norming" (or any other nuisance) effect that might have generally increased (or decreased) the performance in individual trials following group trials, we compared performance in individual trials with no previous group trials (Persons C and F with dominant arm; cf. Table 2) with performance in individual trials that were the first for the nondominant arm but followed group trials with the dominant arm (Person B and E with nondominant arm; cf. Table 2). Note that the nonsignificance of the arm factor in the ANOVA

[6] Please note that in at least one regard, this is a somewhat conservative index of potential motivation gains, because it assumes that any motivation loss (or gain) is due to the altered performance of the less capable member. If it were ever the more capable member who quit first in the dyad, this would be interpreted as a group motivation loss, even if the less capable member was willing and able to exceed his or her individual level of performance. In principle, this oversimplification might have been addressed by noting which person finished first in each dyad trial. Unfortunately, pilot studies suggested that it was sometimes unclear which dyad member actually finished the trial first. For example, sometimes the experimenter observed short nonverbal communications between the participants before one or both hit the bar nearly simultaneously.

reported above suggests that the present contrast could probe the effects of prior group experience without a confound due to arm used. The contrast was not significant, $t(54) = 0.13$. As another, parallel test, we computed the difference between the two individual performance trials (i.e., dominant vs. nondominant arm) for each participant. We then compared these scores for participants who had versus did not have group trials before their first individual trial. Consistent with the preceding analysis, this comparison revealed no significant difference, $t(82) = 0.10$. The same null pattern was obtained when sex of participants was added as second factor in a 2 × 2 ANOVA, all Fs < 1. All these analyses suggest no systematic differences between performance in individual trials before and after the first dyad trials. Moreover, subsequent analyses of the main dependent variable (i.e., CDS scores) revealed no differences between dyads where the weaker coworker performed first individually and dyads where both coworkers performed first in a group, $t(165) = 0.21$, *ns*.

Correction for Fatigue

We corrected for fatigue effects by multiplying all second trials with an arm by the ratio of the first to second individual trial (168/151 = 1.113) and all third trials by the ratio of the first to the third individual trial (168/134 = 1.254). All of the following analyses are based on these corrected data. (There are, of course, other ways one might correct for fatigue effects. For example, one might apply an additive correction, for instance adding to each second performance trial score the mean difference between the first and second performance trial [i.e., 168 − 151 = 17 s] and adding to each third performance trial score the mean difference between the first and third performance trial [i.e., 168 − 134 = 34 s]. However, this correction produced a pattern of results that were basically the same as those obtained with the ratio correction.)

Analyses of Motivation Gains

The weaker member of each dyad was operationally defined as whichever member got the lower corrected score when performing individually. For each dyad trial we computed the ADS, that is, the absolute difference between the weaker and stronger dyad members based on the (corrected) individual trials for the appropriate arm. This ADS score was used as the predictor variable in an analysis where the CDS (i.e., the signed difference between the [corrected] performance of the dyad and the [corrected] individual performance of the weaker person in the dyad) was the criterion variable.

In our primary analyses, like Köhler (1926, 1927), we treated each dyad as an independent data point even though each individual was part of two different dyads. Further, like Ruess (1992), we treated dominant and nondominant arm trials as independent. (For the statistically squeamish, we also did an analysis in which experimental session was used as the unit of analysis, eliminating these problems of potential lack of independence of data points. The results of the latter analysis, which parallels the primary analyses, are reported in Footnote 8.)

The initial inspection of the CDS measure indicated that we indeed found a significant overall motivation gain. The overall mean value of the CDS scores (14.25) was significantly higher than zero, $t(166) = 5.33$, $p < .001$. We then performed a polynomial regression with CDS as criterion and the (centered) ADS as predictor. This analysis indicated that, as with Ruess's (1992) data, the quadratic solution did not explain substantially more variance ($R^2 = 0.04$, adjusted $R^2 = 0.03$) than the simple linear solution ($R^2 = 0.04$, adjusted $R^2 = 0.03$). Note that adding participants' sex and performing arm as additional (categorical) predictors in the regression equation did not improve the level of fit. The best-fitting linear function is plotted in Figure 4. The figure shows that as the discrepancy between dyad members increased, the improve-

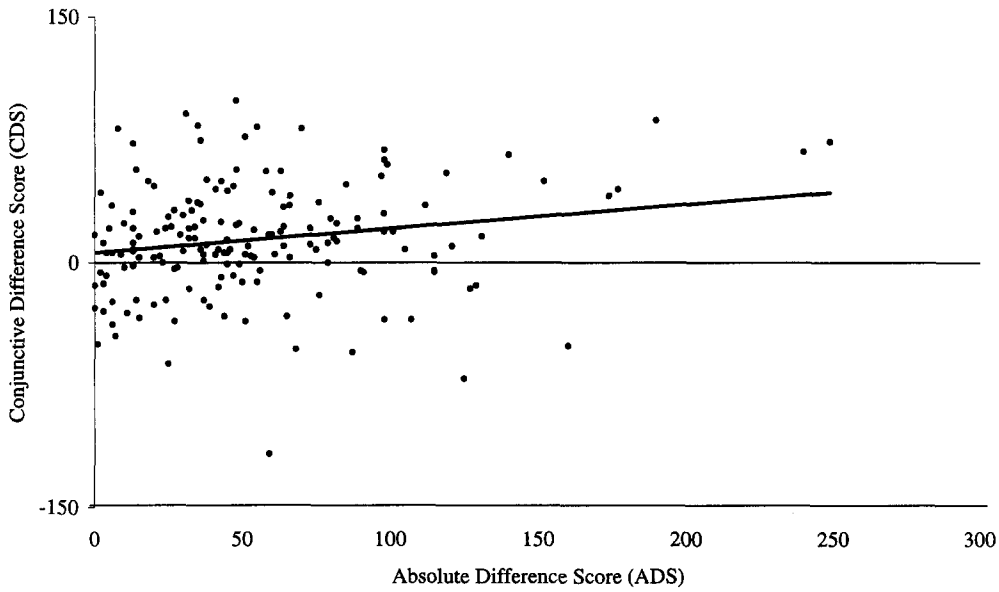

Figure 4. Performance results of Experiment 1.

ment of the dyad relative to its less capable member also tended to increase. The overall simple linear correlation was .19, $N = 167$, $p < .02$.

The preceding analyses suggest the following conclusions: (a) Overall, there is a motivation gain in dyads for our task, (b) this gain effect is moderated by the discrepancy between dyad members performing as individuals, and (c) the function relating discrepancy and motivation gain is positive and linear. In all but one aspect (an inverted-U function), it would appear that Experiment 1 replicated Köhler's (1926, 1927) results. However, before accepting all these conclusions, we must make allowances for any artifactual biases inherent in our analyses.

Correcting for Regression Artifacts

We have made passing reference to the risk of regression-to-the-mean biasing particular statistics of interest. Our current interest was focused on two particular statistics: the mean CDS as an index of overall motivation gain and the (linear) correlation between discrepancy (indexed by the ADS) and relative dyad performance (indexed by the CDS) as an index of the moderation of motivation gain by the discrepancy of member capacities.

Previously, we asserted without proof that (unlike Köhler's ABR) the CDS index is unbiased by regression artifacts. This can be shown logically[7] and confirmed through Monte Carlo simulation (see Appendix B). However, we also mentioned that the second statistic of interest, the correlation between CDS and member discrepancy (indexed by the ADS), was vulnerable to regression artifacts. In particular, the two measures should be positively related, even when there is absolutely no true difference in member effort between performance as an individual and in the dyad (i.e., when any observed differences are attributable to random error). One can reason out why regression effects have this particular qualitative effect on the correlation statistic, but it is even more useful to us in the present context to estimate the magnitude of such regression artifacts, so that we can do standard significance testing (i.e., determine whether any observed correlation departs significantly from what we would expect to see by chance alone).

The details of our simulation are presented in Appendix B. The results indicate that with our sample size, method of dyad composition, and estimated between- and within-subject random error, 95% of the correlations we would expect to see by chance should fall in the interval $.088 \leq r \leq .406$. The observed correlation of .19 is quite near the estimated mean of the sampling distribution (0.250) and clearly not in the $\alpha = .05$ critical region (nor, for that matter, even in the $\alpha = .10$ region). Thus, our data provide no evidence that the motivation gain we observed was moderated by dyad-member ability discrepancy.[8]

Subjective Experience

As noted above, participants subjectively rated their effort as well as perceived importance of own performance, experienced stress, and enjoyment during the task. (Note that the varying degrees of freedom reported for these analyses are due to missing data.) First, we found no significant differences between the ratings in the group and individual trials for the weaker persons in each dyad, as determined by individual trials. Although all their ratings in the group trials were slightly higher (more effort, higher importance to perform well, higher strain, but also more enjoyment) than in their individual trials, none of these differences was significant, $ts < 1.6$. It is noteworthy, however, that the stronger persons in each dyad rated their effort and stress significantly higher in their individual trials ($Ms = 7.7$ and 7.2) than in the group trials ($Ms = 7.4$ and 6.8; $t[153] = 2.43$, $p < .02$), which is not surprising given the conjunctive nature of the collective task condition. The stronger persons only had to work as long as their weaker partner could manage, making group trials less demanding than individual trials. Stronger participants also rated their task enjoyment as significantly higher for individual trials ($M = 4.2$) than for dyad trials ($M = 4.0$), $t(153) = 2.19$, $p < .04$. The stronger coworkers showed no differences between individual and dyad trials on ratings of the perceived importance of performing well at the task, $t < 1$.

However, there might well be reliable individual differences in participants' use of our rating scales. In order to decrease error variance, we decided to analyze relative ratings in addition to absolute ratings. The following analyses were based on the difference scores between ratings in the dyad and on individual trials for each of the rating scales. We focus first on the weaker dyad members.

Analyses of the weaker dyad member's responses. Correlational analyses showed a significant correlation between the ADS and the weaker member's relative rating of subjective effort in the tasks, $r = .325$, $n = 155$, $p < .001$. That is, subjective effort was relatively higher in the dyad as the discrepancy between dyad members grew larger. In addition, the measure of relative subjective effort was correlated with the relative performance gain in the dyads (i.e., the CDS scores), $r = .254$, $n = 154$, $p < .001$. The higher the weaker members rated their effort in the dyad trials

[7] The CDS (conjunctive difference score) is just the difference between the dyad and the poorer of the dyad members' individual scores. The conjunctive nature of the task requires that the dyad score be defined by whichever member quits first in the dyad (i.e., the poorer performance in the dyad). If we can assume that individual levels of performance across testings (individual testing and group testing) differ only due to random error (e.g., error of measurement, chance factors that might affect performance), then there is no reason to expect the lower score obtained by sampling from two individuals' distributions of performance scores on one occasion (e.g., the dyad trial) should differ systematically from the corresponding lower score obtained in a second, independent sampling (e.g., the individual score). Indeed, under the assumptions of our null hypothesis, which occasion we classify as the dyad sampling and which we classify as the individual sampling is arbitrary, so the inherent symmetry of the CDS also reveals that under the null hypothesis of no systematic difference in effort between individual and dyad trials, its expected value should be zero and that this value would not be affected by whether there was relatively more or less random error. Of course, biases due to regression to the mean increase as random error increases.

[8] Another analysis was undertaken in which the experimental session, not the particular dyad testing occasion, was used as the unit of analysis. That is, we computed the mean CDS score and mean ADS scores across all dyads within each session. Although this results in a considerable loss of statistical power, the overall mean of these session CDS scores was significantly greater than zero, session mean CDS = 14.25, $t(13) = 5.76$, $p < .001$. Unsurprisingly, because this type of averaging severely restricts the range of the ADS index of discrepancy, the correlation between session mean ADS and session mean CDS scores was not significant.

(compared with the individual trials), the higher their dyad's (and, given the conjunctive nature of the task, most probably their own) performance in the group (compared with their performances as individuals).

Ratings of relative importance of own contribution by the weaker dyad members were also significantly correlated with performance gains, $r = .214$, $n = 154$, $p < .001$. When weaker participants rated the perceived importance of their contribution higher in the dyad than in the individual trials, the performance was also higher in the dyad. However, the former measure was only weakly related to the objective discrepancy of coworkers (i.e., ADS), $r = .166$, $n = 155$, $p < .04$. It is interesting to note that there was also a significant relation between increase of group performance (i.e., CDS) and the weaker members' relative enjoyment ratings, $r = .170$, $n = 154$, $p < .04$. When the weaker members were working harder, they also reported enjoying the task more.

Weaker members' relative ratings of stress as a measure of negative emotional reactions were not related to relative performance gains in the dyads, $r = .09$, *ns*, nor to the discrepancy of coworkers' capabilities, $r = -.07$, *ns*. This last finding is consistent with Köhler's (1926, 1927) informal observation that the motivating effect of heterogeneous groups was not accompanied by an increase in subjective stress.

In summary, the subjective measures from the weaker dyad member nicely complement the results of the objective performance measures. Increased subjective effort in more discrepant dyads was associated with higher performance in the groups compared with individual trials. At the same time, positive (enjoyment) or negative consequences (stress) of performance in the dyad compared with the individual trials did not covary with the discrepancy of coworkers' capabilities.

Analyses of the stronger dyad member's responses. First, the relative ratings of the stronger dyad members were not related to ratings of the weaker group member, all rs $< .06$. It was not surprising that there was a negative relationship between ADS and subjective effort in the dyad relative to the individual trials, $r = -.216$, $n = 154$, $p < .01$. The greater the discrepancy between members, the lower was these stronger members' self-reported effort in the dyad trials; when discrepancies were large, the stronger member had to quit sooner. The same correlation with ADS was also observed for subjective stress, $r = -.229$, $n = 154$, $p < .002$. These effort and stress ratings were not related to performance increases in the dyad (i.e., CDS), which is unsurprising in light of the fact that it was the weaker, not the stronger, member's performance that determined the dyad's performance.

Discussion

In Experiment 1 we attempted to replicate Köhler's (1926, 1927) findings with a new and safer experimental paradigm. First and most important, there was a significant mean motivation gain across all dyads. Dyads were able to hold their bar up for more than 14 s longer than their less capable member, an increase of about 10% relative to the overall mean of individual performance of the less capable members. However, we did not replicate the moderation of this effect by dyad-member ability discrepancy that Köhler reported. Overall, there appeared to be a tendency for greater motivation gains as group members differed more in ability, but the magnitude of this correlation was also shown to fall well within the limits of what could plausibly be expected due to chance. In this regard, we replicated Ruess (1992); he too obtained a linear correlation between discrepancy and motivation gain, but it too could be plausibly attributed to chance. The clear implication of the results, then, is that for the present experimental task, the motivating effects of working in the dyad on the less capable member are fairly consistent and uniform for dyads with members that are equal, moderately unequal, or extremely unequal in ability, as indexed by individual performance. (Henceforth in this article, when we refer to "the Köhler effect," we will mean the overall motivation gain Köhler observed and not the moderation by member discrepancy in ability he also reported.)

In other regards, our data extend and qualify Köhler's findings. As the structure of a conjunctive task would suggest, it was the less capable member of the dyad who was critical for producing the observed motivation gain. When the less capable members saw their contribution as indispensable to the group, their dyads performed better; no such relationship was observed for the more capable members. As the discrepancy between partners increased, weaker members reported exerting more effort, whereas stronger members reported exerting less effort and experiencing less strain.

Experiment 1 confirmed the utility of our modification of the Ruess (1992) task and procedure as a paradigm for obtaining Köhler's overall motivation gain effect. In Experiment 2, we again used that paradigm to competitively test two explanations for these motivation gains.

Experiment 2

Several alternative explanations have been offered for the Köhler effect (see Stroebe et al., 1996; Witte, 1989). Here we focused on the two explanations that seem the most plausible in the light of all available data.

Goal Comparison

Stroebe, Diehl, Abakoumkin, and Arnscheid (1990) suggested that when there is no clear standard of good performance, group members engage in social comparison of one another's level of performance to decide on reasonable performance goals. They go on to suggest that when task accomplishment is unimportant or not valued by group members, there will be a downward bias in this social comparison process—the more capable members will adjust their goals downward to be more in line with the less capable members (also see Paulus & Dzindolet, 1993). Conversely, if task accomplishment is important or valued by group members, the bias should be in the opposite direction—that is, those performing less well should set goals closer to the performance levels of the most capable group members. Stroebe et al. (1990) plausibly argued that the weightlifting task that Köhler (1926, 1927) had his rowing team members perform was an important training activity for the club, and thus the less capable members should have set higher goals in the dyad conditions, producing genuine motivation gains when working at the conjunctive group task. Köhler's curvilinear function is explained by suggesting that no adjustment of goals is necessary when dyad members perform at the same level (very low discrepancy) and that the weaker member will not accept his partner's level of performance as a reachable goal when there is a

very high discrepancy in abilities. If participants were not well informed about one another's capabilities, as in Experiment 1, this social comparison process might not be moderated by actual discrepancies in ability.

Perceived Indispensability

Kerr (1990) suggested that Köhler's (1926, 1927) group motivation gain effect might be explained using the same family of models most frequently used to account for group motivation loss effects (cf. Karau & Williams, 1993, in press; Kerr, 1983; Shepperd, 1993) and for at least one apparent motivation gain effect (viz. Williams & Karau's, 1991, social compensation effect), namely, Instrumentality × Value (I × V) models (cf. Vroom, 1964; Stroebe & Frey, 1982). These models hold that one's choice of effort level is governed by how instrumental that level of effort is for achieving an outcome, weighted by the value placed on that outcome (summed, typically, across all possible outcomes).[9] So, for example, one would be more likely to choose a low level of effort if one perceived one's contributions as highly dispensable for the group (i.e., low instrumentality between personal effort and group success). Such dispensability-driven free riding has been well documented in the group motivation-loss literature (Harkins & Petty, 1982; Kerr, 1983; Kerr & Bruun, 1983). Particularly relevant in this context is Kerr and Bruun's (1983) finding that a low-ability group member will work harder than a high-ability member at a nonyoked, conjunctive task. The Köhler effect suggests that this higher level of performance can even exceed the level of performance of the individual low-ability member.

This explanation suggests that it is the conjunctive nature of Köhler's task that makes the performance of the less capable member crucial for the group's success. At such a task, as soon as one member quits, the other must quit very soon as well. Hence, the less capable member is likely to see his or her efforts as particularly indispensable for the group's success. However, we suspected that the group's success–failure is not the only (or, perhaps, even the most important) effort-contingent outcome for the less capable dyad member in Köhler's paradigm. There may also be highly salient interpersonal evaluations at stake. When one person in the dyad quits at Köhler's task, he or she (a) compels her partner to quit before the partner wants to quit and (b) is likely to be seen as personally responsible for both the partner's and the group's not doing any better.[10] Both of these judgments are likely to be aversive and stigmatizing, especially for people like Köhler's rowers, for whom doing well at the task and the esteem of fellow club members were important.

Thus, if the I × V explanation is valid, it is possible that the salient, valued outcomes that underlie the Köhler effect have to do primarily with (a) task success or failure, (b) positive impression management, or (c) some mix of both. It was not our purpose in this experiment to settle this particular theoretical question. Rather, we wanted to test the I × V model's more basic assumption that whatever the key outcome(s), high effort by the weaker dyad member is very instrumental for producing that outcome(s) in the Köhler paradigm. Moreover, we wanted to contrast this explanatory mechanism with the different, goal-comparison mechanism of Stroebe et al.'s (1990) model (see also Stroebe et al., 1996).

How does the I × V explanation account for Köhler's nonlinear function? When dyad members' abilities are nearly equal, they both want to quit at the same time, so neither is seen as "holding back" the other or the group. (Even a little coordination loss would account for the small process loss observed by Köhler when coworkers are equal in capability.) When the discrepancy is very large, it should become very apparent to the less capable member (from one another's apparent level of fatigue or perhaps from foreknowledge of one's partner's strength) that there is no way that he or she could match the performance of the much more capable partner, and so he or she gives up. But when the difference in ability is moderate, we suspect that the less capable member can entertain hopes of matching his stronger partner, lift for lift, and he or she should therefore persist as long as he or she possibly can (to prolong group performance and to avoid the stigmas marking the person that quits first).

How could this explanation account for the results of Experiment 1, that is, motivation gain unmoderated by the discrepancy in members' abilities? If, as we argued in our General Discussion, there was considerable ambiguity about members' relative capabilities, as in Experiment 1, the conjunctive nature of the task could make both dyad members willing to work harder, regardless of any discrepancy in capability. However, the less capable member would be expected to reach his or her genuine performance limit (i.e., his or her true maximum capacity) and quit sooner.

A Competitive Test

The primary purpose of Experiment 2 was to contrast the latter, I × V/dispensability explanation against Stroebe et al.'s (1990) goal comparison explanation. To this end, we examined the performance of dyads compared with individual controls under both conjunctive and additive task demands. All participants first per-

[9] There are many theoretical variations on this basic theme. Some models (e.g., Vroom, 1964) have differentiated the contingency between an action and outcome, on the one hand, from the contingency between the outcome and a salient reward or punishment, on the other. Other models (e.g., Karau & Williams, 1993, in press) distinguished between the expectancy that one can achieve a certain level of performance (much akin to perceived self-efficacy; Bandura, 1986) and the expectancy that this performance will produce a particular valued (or disvalued) outcome. And, these different models might use somewhat different terminology than we have used here (e.g., some might use the term "expectancy" where we use "instrumentality"). Although such distinctions may be useful for certain purposes, we believe a rather generic I × V model, as is outlined in the text, is sufficient for our present purposes.

[10] There are other interpersonal outcomes that result from a weaker member's quitting first. In particular, in so doing, (a) he or she concedes that he or she cannot keep up with the partner, and (b) as such, he or she is identified—to the partner, the experimenter, and any observers—as the less capable member of the group. Note, however, that these outcomes are not distinctive to the yoked-conjunctive tasks used by Köhler (1926, 1927) and us (in Experiment 1); they should also arise for any task where comparison of dyad member performance is possible (including many additive or disjunctive tasks). Moreover, it may be precisely to avoid these outcomes that a weaker member might set a goal closer to his stronger partner's level of performance. If this were the case, the goal comparison model could be considered to be just a variant of the I × V model. However, here we wanted to explore whether it is the criticality of the weaker member's effort that produces the Köhler effect, as the I × V explanation suggests, or the possibility of interpersonal comparison, as Stroebe et al.'s (1990) goal comparison model posits.

formed an individual trial and then a group trial under one of these two task demands. The *conjunctive task* version was very similar to the task requirements of Experiment 1. One potentially significant difference was that in Experiment 2 dyad members were not physically yoked by holding a single bar. Rather, each of the two dyad members held their own bar. As in Experiment 1, however, the trial was over when either one of the dyad members quit the task and hit the flexiglas trip bar. In the *additive task* demand condition, a dyad trial was not over when one dyad member quit; the other dyad member could continue as long as possible and thereby earn more points for the team.

According to the I × V/dispensability explanation, the Köhler (1926, 1927) motivation gain should only occur in the present conjunctive condition, because it is only in this condition that the weaker member is indispensable for dyadic performance, sets limits on the stronger member's performance, and may be seen as responsible for group or partner failure. However, the processes of social comparison and goal setting can operate equally well under both additive or conjunctive task demands. Hence, the goal comparison explanation predicts that motivation gains by the weaker dyad member should occur in both the conjunctive and additive conditions. Moreover, because this explanation holds that the weaker member tries to match the performance of the stronger member, it also predicts that motivation gains should increase as the discrepancy in member abilities increases. Of course, it is also possible that both processes contribute to the Köhler motivation gain. In this case, we would expect to see a weaker motivation gain under additive task conditions (where only the goal-comparison process could operate) than under conjunctive task conditions (where both processes could operate).

In addition to the additive and conjunctive conditions, some participants in Experiment 2 performed in an individual control condition. Here there were no group trials, but repeated individual trials. This condition was included for two reasons: first, to enable proper estimation of fatigue effects, and second, to provide an empirical basis for estimating regression effects.

Method

Participants and Design

Seventy-two female undergraduate students at Michigan State University participated in partial fulfillment of a course requirement. Because no sex effects emerged in Experiment 1, and because there were considerably more women in the potential participant pool, all participants were women. There were two pairs of experimenters, one man and one woman. To check to see if sex of experimenter might moderate our results, this factor was included in the experimental design.

Hence, the experiment used a 3 (task demands: additive, conjunctive, individual controls) × 2 (sex of experimenter) × 2 (performing arm) × 2 (trials) design, with the two last factors' being within-subject variables. Sessions were conducted with 4 female participants. Each session was assigned randomly to one of the six between-subject conditions. Unlike Experiment 1, order of performing arm was counterbalanced, and the order of individual and group trials was held constant (individual trials always preceded dyad trials).

Experimental Task

There were a few differences between the experimental tasks of Experiments 1 and 2. In Experiment 2, dyad trials were not conducted with a metal bar twice as long and as heavy as the bar used by individuals, but instead each dyad member held up bars of the same size and weight (viz. .69 kg) as those used in the individual trials. In addition, each dyad member had her own flexiglas "trip bar," balanced between two poles. These modifications were necessary because in the additive group condition it was possible for one participant to continue after the other had stopped. These changes also made it possible to determine whether physical yoking of dyad members (as in Experiment 1 and Köhler's experiments) was necessary to produce the Köhler motivation gain effect.

Procedure

With but a few exceptions, the procedures were identical to those used in Experiment 1. However, the instructions were adapted for four-person (rather than six-person) experimental sessions.

In the individual condition it was explained,

> On some trials, some of you will be performing this task in Room 1, in other trials, some of you will be performing in Room 2. Your score on either room is just the total number of seconds between the start of the trial and when your arm hits a flexiglas bar . . . we will sum up all the scores on all the trials to come up with a Total Team Score.

In the additive condition it was explained,

> On some trials, you will be performing this task as an individual in Room 1. On other trials you will be performing it with another person in Room 2. We will call these two-person trials in Room 2 the dyad trials. The score in the dyad trials is the total number of seconds that each person keep their metal bar above their flexiglas bar. The trial is not over until both persons' arms have hit their flexiglas bars . . . we will sum up all the scores on all the trials today to come up with a Total Team Score.

In the conjunctive condition it was explained,

> On some trials, you will be performing this task as an individual in Room 1. On other trials you will be performing it with another person in Room 2. We will call these two-person trials in Room 2 the dyad trials. The score in the dyad trials is the total number of seconds between the start of the trial and when either person's arm hits a flexiglas bar. That is, when either person's arm hits her flexiglas bar, the trial is over for the dyad . . . we will sum up all the scores on all the trials today to come up with a Total Team Score.

The incentives offered were the same as in Experiment 1. The experimenters also stressed that participants should perform as well as they could, but should lower their arm if it became too tired or uncomfortable.

Performance trials followed a fixed schedule (see Table 3). Which experimenter worked in which room was alternated across experimental sessions. Participants were assigned randomly to the letter roles of the schedule. In Experiment 2, order of performing arm was counterbalanced and order of individual and group trials held constant. The first trial for each participant was an individual trial, half of participants performing with the dominant arm and the other half with the nondominant arm. Participants' second trial was performed with the same arm as the first trial (see Table 3). In the individual control condition, each participant's second performance trial was again an individual trial. The third and fourth performance trials for each participant repeated the conditions of the first two but were performed with the other arm. As in Experiment 1, the observing audience could not talk to the performer(s) during the trial. And once again, after each trial, participants completed a short questionnaire measuring their subjective experiences (subjective effort, perceived importance of their own contribution, experienced stress and fun) during the trial.

Table 3
Trial Schedule of Experiment 2: Additive and Conjunctive Condition

	Room 1		Room 2		
Trial–arm	Performing	Audience	Performing	Audience	Resting
1–Dom.	A	D	B	C	
2–Nondom.	D	A	C	B	
3–Dom.	A, B	C			D
4–Nondom.	C, D	A			B
5–Nondom.	B	D	A	C	
6–Dom.	C	A	D	B	
7–Nondom.	A, B	D			C
8–Dom.	C, D	B			A

Note. Dom. = dominant arm; Nondom. = nondominant arm. Letters A–D stand for the 4 participants in each session. In the individual condition, in Trials 3 and 7 Participants A and B were assigned to different rooms performing alone with one of the resting participants as audience. The same holds for Trials 4 and 8 and Participants C and D, respectively.

After finishing a session, participants were debriefed and thanked. At the end of the experiment, one group was picked at random and, as promised, received a monetary reward for their performance.

Results

Performance in the Individual Condition

First, we analyzed the performance times in the individual condition in order to ascertain and later adjust for fatigue effects, and to test whether sex of experimenter or order of performing arms (dominant first, nondominant second vs. the reverse) affected the performance across the four trials per participant. Recall that each of these participants performed with one arm on the first two trials with a rest trial between performance trials; we refer to this set as Block 1. On the two following trials, they then performed with the other arm, again with a rest trial in between; we refer to this set as Block 2. A 2 (sex of experimenter) × 2 (order of performing arms) × 2 (trial block) × 2 (Repetition: first vs. second performance trial per block) multivariate analysis of variance (MANOVA) with repeated measures on the last two factors revealed two significant effects. The first was a main effect of the repetition factor, $F(1, 16) = 23.07$, $p < .001$, reflecting a general fatigue effect. Performance times on the first trial in each block ($M = 145.0$ s) were significantly longer than on the second ($M = 116.5$ s). There were no interactions with the block factor or the order factor, all other Fs < 1, showing that the fatigue effect was similar for dominant and nondominant arms and independent of the order of arm used. Based on these results we computed the ratio of performance on the first to the second individual trial (1.245) as a fatigue correction factor, similar to the fatigue correction applied in Experiment 1.

There was no significant difference between the two trial blocks performed with different arms, $F < 1$. Performance with one arm in the first trial block did not significantly affect the performance of the other arm in the following trial block. However, there was a significant interaction effect of the order factor and the trial block factor, $F(1, 16) = 7.61$, $p < .02$. Recall that in the first trial block, half of the participants started with their dominant arm and half started with their nondominant arm. Participants who started with the dominant arm performed higher in the first trial block ($M = 137.2$ s) than in the second trial block ($M = 130.3$ s). The reverse pattern was found for participants who started with the nondominant arm ($M = 119.0$ s vs. 136.5 s, respectively). In other words, in both order conditions performance with the dominant arm was significantly higher than performance with the nondominant arm. (Recall that Experiment 1 participants also performed [somewhat] better with their dominant arm, but in that study, the difference was not significant.) This arm effect was not moderated by any other factor in this analysis, all Fs < 1.2. Finally, sex of the experimenter had no significant effect on the performance times of the female participants in the individual condition, $F < 1$.

Overall Motivation Gains

The index of group motivation gains in the present conjunctive condition was the CDS, the difference between the fatigue-corrected dyad performance and the worse of the two dyad members' individual performances. In the additive condition, the dyad performance score was defined by the fatigue-corrected score of the first dyad member to quit. Finally, we were able to define a "pseudo-dyad" CDS score in our individual control condition. It was just the difference of the lower corrected score of the two individuals on their second testing to the lower score of those same two individuals on their initial testing. In Experiment 1 we had to rely solely on simulations to determine whether certain statistics of interest (e.g., the correlation between member discrepancy and motivation gain) were statistically significant. In Experiment 2, we were also able to use the pseudo-dyad CDS scores in the individual control condition to provide an empirical "chance" baseline with which to compare performance in the conjunctive and additive conditions. And, as in Experiment 1, trials with the dominant and the nondominant arm were treated as independent cases.

The mean CDS scores differed significantly between our three experimental conditions, $F(2, 66) = 8.52$, $p < .002$. Motivation gains in the conjunctive condition (mean CDS = 45.70; $SD = 33.03$) were significantly greater than the chance baseline of zero, $t(26) = 7.19$, $p < .001$, as well as the corresponding means in the additive (mean CDS = 11.90, $SD = 46.60$), $t(66) = 3.37$, $p < .002$, and the individual (mean pseudo-CDS = 7.54, $SD = 18.38$), $t(66) = 3.61$, $p < .007$, conditions. The latter neither differed from one another, $t(66) = 0.40$, *ns*, nor from the chance baseline of zero, $t(22) = 1.22$, *ns*, for the additive condition; $t(18) = 1.79$, $p > .09$, in the individual control condition.

Thus, it was only in the conjunctive condition that significant overall motivation gains were observed. Dyads in this condition performed about 35% longer than the individual performance of the dyad's less capable partner. This pattern of results is inconsistent with the goal comparison explanation, which predicts a motivation gain in both the conjunctive and the additive conditions, but it is fully consistent with the I × V/dispensability explanation, which predicts a motivation gain only in the conjunctive condition. A parallel analysis that considered the sex of the experimenter indicated that the preceding findings were not moderated by this factor.

Relationship Between Member Ability Discrepancy and Motivation Gain

Next, within each experimental condition we computed the correlation between the ADS measure of member ability discrepancy[11] and the CDS measure of motivation gain. The first finding of interest was that the correlation in the individual control condition was substantial and significant, $r = .54$, $n = 19$, $p < .02$. This finding corroborates the results of the Monte Carlo simulation in Experiment 1 (see Appendix B). Regression artifacts alone will produce an apparent positive association between member discrepancy and indices of group motivation gain that focus on the poorer performance in the dyad. Another simulation, similar to the one outlined in Appendix B (with 500 replications and the appropriate sample size, $n = 19$), indicated that the observed correlation of 0.54, although substantial, was within the limits of chance (the mean estimated correlation was 0.289 with a 95% confidence interval of $-.147 \leq r \leq .661$; the estimated p value for the observed $r = .54$ was .26, two-tailed).

In the conjunctive condition, the correlation was also substantial, $r = .42$, and, compared with the (inappropriate) chance baseline of $r = 0$, significant $t(26) = 2.31$, $p < .03$, two-tailed test. However, it was neither significantly different from the corresponding correlation $r = .54$ in the individual control condition ($z = 0.49$, *ns*), nor did it fall outside the chance limits defined by simulation (for $n = 27$, the estimated mean correlation was 0.283, and the 95% confidence interval was $-.103 \leq r \leq .621$). Thus, as in Experiment 1, we found no evidence for the discrepancy of member abilities moderating the observed motivation gain under our conjunctive task conditions. (For completeness sake, we also looked for a nonlinear [i.e., quadratic] component for the function relating [centered] discrepancy and motivation gains in the conjunctive condition. Higher-order solutions did not explain more variance [quadratic model: $R^2 = 0.19$; adjusted $R^2 = 0.12$], compared with the simple linear model [$R^2 = 0.17$; adjusted $R^2 = 0.14$].)

Finally, we computed the same correlation in the additive condition. As noted above, the goal comparison explanation predicts that motivation gains should increase as the discrepancy between dyad members increases. However, the observed correlation was actually negative ($r = -.09$), although not "significantly" so, relative to the (inappropriate) $r = 0$ baseline ($n = 23$, $t[22] = 0.41$, *ns*). This result does not simply reflect a pronounced nonlinear (e.g., quadratic) effect resulting in a near-zero correlation. Regression analyses produced no linear or nonlinear (e.g., quadratic) trends in the additive condition. Another simulation with the appropriate sample size ($n = 23$) indicated that the observed correlation was well within the actual limits of chance (estimated mean correlation = .206, with a 95% confidence interval of $-.354 \leq r \leq .620$).

Performance of the Stronger Coworker

The analyses of the stronger coworkers' performance was restricted to the additive and individual conditions, because in the conjunctive condition it was identical with the performance of the weaker coworker due to the nature of the task. As one would expect from regression to the mean, the dyad or pseudo-dyad partner who did better on the first trial tended, on average, to do 7.44 seconds (corrected for fatigue) less well on her second trial; however, this trend was not significant, $t(41) = 1.29$. More important, this decline in performance did not differ significantly between the additive ($M = -8.15$) and individual control ($M = -6.59$) conditions ($t < 0.2$). Thus, working together with a weaker coworker under additive task demands resulted in neither distinctive motivation losses (relative to individual controls) nor in distinctive motivation enhancements (e.g., due to the mere presence of a coactor, competitive pressures, or the desire to socially compensate for the weaker partner) by the stronger coworker.

Subjective Ratings of the Weaker Coworkers

In addition to the objective performance measures we again analyzed the participants' ratings of effort, perceived importance of own performance, enjoyment of the task, and stress, collected after each trial. First, we focused on the ratings of the weaker coworker in each dyad (including the individual condition, where a "dyad" consisted of the two persons who performed at the same time, although in different rooms). Overall, these participants rated their subjective effort generally higher in the second ($M = 6.9$) than in the first trial ($M = 6.6$), $t(68) = 3.08$, $p < .01$. A similar result was found for ratings of stress ($M = 5.3$ and 6.4, respectively), $t(68) = 6.14$, $p < .001$. Ratings of the other two variables did not differ on average between the two trials, $ts < 1$.

To permit a more detailed analysis of differences between the performance conditions, we created relative ratings, as in Experiment 1, by computing difference scores of the two consecutive trials for each person (T2 − T1), so that positive scores indicate higher ratings in the second trial (i.e., the group trial in the additive and conjunctive condition). Comparing the effort ratings of the weaker dyad member in the conjunctive and the additive conditions revealed a greater increase in subjective effort in the conjunctive condition ($M = 0.44$) compared with the additive condition ($M = 0.35$). Although this difference was not significant ($t < 1$), it is consistent with the objective performance data. Indeed, increases in subjective effort correlated significantly with increases in performance times, $r = .248$, $n = 69$, $p < .05$.

Moreover, comparing the relative ratings of enjoyment and perceived importance of one's own contribution between the conjunctive and additive condition revealed significant differences (see Table 4 for details of this contrast, as well as for the individual condition data). In contrast to the additive condition in which weaker persons perceived their contribution as less important on the second trial compared with the first individual trial ($M = -0.35$), in the conjunctive condition weaker members rated their contribution in the group trials as more important ($M = 0.22$),

---

[11] As in Experiment 1, we used the ADS (absolute difference in dyad members' individual performance scores) as an index of ability discrepancy. In Experiment 2, some of these difference scores were quite extreme, suggesting that some of the very poor initial performances may well have been due to irrelevant factors (distraction, unfamiliarity with the task, etc.). To eliminate any possible biasing effects of these extreme scores, we excluded three cases where the observed ADS scores were more than two standard deviations above the overall mean. It might be noteworthy that it was only in these three cases that the individual performance of the weaker person was less than 1 min. There was one such case in each of the three experimental conditions (no such case was found in Experiment 1).

Table 4
Mean Differences of Ratings After the Second and First Performance Trials (T2 − T1)

Condition	Effort	Importance	Enjoyment	Stress
		Weaker dyad member		
Conjunctive	0.44	0.22	0.56	1.33
Additive	0.35	−0.35	−0.22	1.09
Individual	0.21	0.32	−0.37	0.84
		Stronger dyad member		
Conjunctive	0.96	0.93	0.59	1.48
Additive	0	0.09	−0.04	1.09
Individual	−0.05	−0.21	−0.11	1.94

$t(48) = 2.05$, $p < .05$. Similarly, in the additive condition the weaker coworkers rated the group trials as less enjoyable than the individual trials ($M = -0.22$), presumably because of general fatigue (note the comparable difference score in the individual control condition, see Table 4). However, in the conjunctive condition the weaker coworkers rated the group trials as more enjoyable than the individual trials ($M = 0.56$), and this difference was significantly larger than in the additive condition, $t(48) = 2.71$, $p < 0.01$. This last result is especially interesting because increases in enjoyment ratings between first and second trial correlated with performance increases between the individual and group (or psuedo-group) trials (in all dyads), $r = .314$, $n = 69$, $p < .01$. On the other hand, the ratings of stress in the conjunctive ($M = 1.33$) and additive condition ($M = 1.09$) did not differ significantly.

Overall, these analyses suggest that the conjunctive and additive group trials were experienced differently by the less capable member, with dyad conditions' enhancing felt enjoyment more in the conjunctive condition than in the additive (or the individual control, cf. Table 4) condition and that this different experience was linked to performance gains in the group trials.

Subjective Ratings of the Stronger Coworker

We also analyzed the ratings of the stronger coworkers in each dyad and pseudo-dyad in order to explore the motivational consequences of the different performance conditions. For example, the stronger dyad partner might enjoy working with the weaker partner less than in the individual pseudo-dyads (cf. Experiment 1). On the other hand, there might be certain working conditions under which the stronger coworker might enjoy working in a group more than alone. Remember that no direct comparison of the objective performance data of the stronger coworker was possible due to the demands of the conjunctive task.

First of all, the absolute rating patterns were similar to those of the weaker coworkers. Subjective effort and stress in all three conditions was generally rated higher in the second trial than in the first trial (M_{effort} = 7.1 vs. 6.7), $t(68) = 2.81$, $p < .01$; (M_{stress} = 6.7 vs. 5.2), $t(68) = 7.14$, $p < .001$. Second, the perceived importance of own contribution was generally perceived higher in the second than in the first trial ($M_{importance}$ = 7.1 vs. 6.8), $t(68) = 2.70$, $p < .01$. Finally, there was no overall trial difference for the enjoyment ratings in Experiment 2, $t < 1.3$.

We again compared the difference scores (T2 − T1) of the stronger members' ratings in the additive and conjunctive conditions in order to investigate whether these overall trial differences were moderated by the different group conditions. In the conjunctive group trials, the stronger dyad members reported a greater increase in effort ($M = 0.96$) than in the additive group trials ($M = 0$), $t(48) = 3.59$, $p < .01$ (see Table 4 for the individual condition). A similar pattern was found comparing perceived importance of the own contribution for the conjunctive ($M = 0.93$) and additive condition ($M = 0.09$), $t(48) = 2.91$, $p < .01$, and, to a lesser extent, for the ratings of enjoyment during the trials ($M = 0.59$ vs. -0.04), $t(48) = 1.79$, $p < .09$. The difference between the conjunctive and additive conditions in relative ratings of stress did not differ significantly, $t < 1$.

These results suggest that enhanced perception of enjoyment, perception of the importance of one's own contribution, and perception of effort under conjunctive task demands in Experiment 2 were not restricted to the weaker dyad member, but also seemed to characterize the stronger group members. As a consequence, it might be reasonable to assume that the stronger coworkers were also more motivated in the group compared with the first individual trials, although (because we have no performance score for the more capable member in the conjunctive condition) this could not be detected in the objective performance data.

Although there was no correlation between the relative effort ratings of the weaker and the stronger coworker in the additive condition ($r = .06$), this correlation was significant in the conjunctive condition ($r = .47$, $p < .02$). One possible explanation for this last finding is that the enhanced effort of the weaker member "pushed" the stronger member to maintain or even increase her performance, so that she (the stronger member) would not end up being responsible for the dyad quitting.

Discussion

Experiment 2 extends our understanding of the Köhler (1926, 1927) effect both empirically and theoretically. The theoretical insights are particularly interesting. As one would expect from the I × V/dispensability explanation, it was only when the task demands were conjunctive that significant levels of motivation gain were observed. Further, as one would expect from this explanation, the weaker dyad member felt more indispensable under conjunctive task demands than in the other conditions. Contrary to the social comparison–goal setting explanation, there was no motivation gain observed under additive task conditions, nor was the level of performance of the less capable member significantly moderated by the discrepancy between dyad member capabilities. The failure to observe motivation gains under additive task conditions also suggests that the effect cannot be attributed to the value placed on certain "social-comparison" outcomes (see Footnote 10). For example, if weaker members worked harder simply to avoid being identified as the less capable member of the group, we should have observed a motivation gain in the additive condition. Finally, a possible (although implausible; see Latané, 1981; Kerr & Yukelson, 1983) alternative explanation for the motivation gain in Experiment 1 was that participants felt greater evaluation apprehension in dyads simply because the dyad partner represented one more observer and evaluator. However, this explanation is contradicted by the absence of any detectable motivation gain in the

present additive condition. In summary, based on available evidence, the I × V/dispensability explanation appears to offer the best explanation for the Köhler motivation gain effects we have obtained.

Empirically, we replicated and extended the basic findings of Experiment 1, observing significant motivation gains for our conjunctive task that were not moderated significantly by the discrepancy in dyad members' capabilities. It is interesting to note that Stroebe et al. (1996) took their failure to observe moderation of the motivation gain they observed in their crank-turning experiments (e.g., Experiments 2–4) as a failure to replicate the Köhler effect. However, our results suggest that the mechanism underlying the motivation gains observed by Köhler may, at least under some circumstances, not be sensitive to the discrepancy of group members' capabilities. Thus, perhaps the motivation gains observed by Stroebe et al. (1996), at least in part, stemmed from this same dispensability mechanism, and not, as they speculated, primarily from the motivating effects of intragroup competition. On the other hand, a couple of aspects of Stroebe et al.'s (1996) results suggest that they were correct—that is, that a different mechanism is at work with their crank-turning task. First, they found some evidence of motivation gains by the stronger member of the dyad, especially when members were similar in ability. We found no such evidence (in the additive conditions of Experiment 2); further, such an effect is precluded by the conjunctive nature of Köhler's and the present conjunctive group task. Second, in one study (Experiment 3), they found the same degree of motivation gain under additive and conjunctive task demands, unlike our findings in Experiment 2.

Now that we have successfully replicated and begun to explain the Köhler effect, we can better support the argument that this is a distinctive group motivation gain phenomenon and not simply a manifestation of some other such phenomenon (also see Witte, 1989). Our results show that the Köhler effect is not just social facilitation (Experiment 2 shows that it is not the number of observers present that is crucial), not just attributable to the effects of intermember competition (Experiment 2's additive condition contradicts this possibility), clearly not dependent upon mixed-sex group composition (nor, it seems, on particular levels of group ability composition), clearly not the result of explicit goals that are preset and difficult (no such goals were set in either Köhler's [1926, 1927] or our studies), and clearly not a form of social compensation (because the Köhler effect arises from the weaker member's working harder, not the stronger member's compensating for an ostensibly weaker member, as in social compensation).

The few differences between Experiments 1 and 2 help to extend the domain to which the Köhler motivation-gain effect can be generalized. For example, we find the effect both when individual trials always precede dyad trials (in Experiment 2) and when the order of individual and dyad trials is varied in a systematic fashion (Experiment 1, see Table 2). In Experiment 1 the experimenters were always the same sex as the participants; in Experiment 2, we found that the Köhler effect was not moderated by the sex of the experimenter (at least for female participants). Most interesting, we found the Köhler effect in Experiment 2 with dyad members' holding separate bars rather than the single bar used in Experiment 1. Thus, physically "yoking" group members (which was also a feature of Köhler's [1926, 1927] original lifting paradigm) seems not to be necessary to produce the overall Köhler effect. This finding is not inimical to the I × V/dispensability explanation. Although such physical yoking may be useful for identifying which member is approaching his or her performance limit first and is, hence, particularly indispensable to the group, with continuous veridical performance feedback from all group members during each performance trial, yoking is not essential for correctly perceiving relative member indispensability.

The subjective measures again complemented the results for the objective performance measure. Participants seemed aware of how hard they were working; subjective effort increases were significantly correlated with objective performance gains. Under conjunctive task conditions, where a genuine motivation gain occurred, the weaker members saw their input as more important to their dyad and thought they worked harder in the dyads (compared with working as individuals), whereas the opposite was true for the weaker member of dyads working under additive task demands. It is interesting that the greater perceived indispensability and motivation gain of the weaker members in the conjunctive condition were not accompanied by greater perceived stress. To the contrary, these participants reported enjoying working at the task in dyads (relative to their individual performance) significantly more than their counterparts in the additive condition. Moreover, in Experiment 2 we found rather similar results for the stronger dyad members in the conjunctive dyads. These participants also felt relatively more important to their group, felt like they were exerting greater effort in their group, and enjoyed group work more without experiencing greater stress than their counterparts in the additive condition. These last findings might seem paradoxical; the stronger members in the additive condition could actually contribute directly to their dyad's total score and potential earnings, whereas the stronger members in the conjunctive condition could not. We speculated that under conjunctive task demands, the weaker member "pushes" the stronger member, requiring him or her to maintain a relatively high level of performance, lest he or she be outperformed by a less capable partner and become the person who limits the group's performance. So, the conditions under which the Köhler effect obtains have the happy conjunction of the group objectively doing better than one would expect (from individual performances), along with both dyad members' seeing themselves as more important, working harder, and yet enjoying themselves more. These conditions appear to challenge group members to do better, and they seem to enjoy trying to meet this challenge.

Of course, other explanations (e.g., justification of effort; see Aronson & Mills, 1959) cannot be ruled out with our data. Another interesting possibility is that weaker (and perhaps stronger) members anticipate doing better during the dyad trials, and the consequent enhanced mood might itself enhance one's willingness to continue working (e.g., a more positive mood might be taken as information that one is experiencing little difficulty persisting at the task; Martin, Ward, Achee, & Wyer, 1993; Schwarz & Clore, 1996). Of course, in our paradigm, there is no way of determining whether enhanced mood is a consequence or a cause of better performance. Clearly, the effects of dyad member affect in the Köhler paradigm have implications both for application to real work groups and for better understanding of the Köhler effect itself.

General Discussion

By way of summary, the studies reported here have accomplished several things. First, a practical and safe experimental task and paradigm has been developed within which the motivation gain effect originally reported by Köhler (1926, 1927) can be observed reliably. Second, it has been shown that, contrary to Köhler's findings, this motivation gain effect does not seem to be moderated by the discrepancy in member capabilities, at least within the present paradigm. Third, consistent with Köhler's informal observation, it has been shown that this motivation gain is accompanied by increased enjoyment of the task and without increased experienced stress. Fourth, it has been shown that the Köhler effect is most likely the result of task demands making the less capable member of the group feel particularly indispensable for group success (and, possibly, desirous of avoiding certain stigmas that could accompany being the first to quit at a conjunctive task). The available evidence makes it unlikely that the effect is due to weaker members' trying to match the performance levels of stronger members through a process of social comparison.

The new task paradigm was developed by incorporating as many features of Köhler's original task paradigm as possible. Having successfully produced the effect, one strategy for further analyzing it is a "subtractive" approach. This involves systematically removing various features to see if the effect is eliminated or attenuated, thereby identifying potential necessary or facilitative conditions. The work completed thus far has effectively applied this strategy. We now know, for example, that a number of features present in Köhler's work, but absent in our paradigm, are not absolutely necessary for the overall motivation gain effect. These include (a) verbal encouragement or exhortation by others in the performance dyad or a more superordinate group (the rowing club in Köhler's study; the full session group in our studies), (b) long-term or continuing membership in that superordinate group, (c) high intrinsic interest in the performance task, and (d) physical yoking of members of the performance group. On the other hand, we also know that it is probably necessary to use a task that makes the least capable member feel indispensable for group success. There remain many other features that can and should be examined using this subtractive method. For example,

1. Is it sufficient that group members simply understand the conjunctive nature of the task, or must they have the kind of continuous, on-line feedback of members' performance that alerts the less capable member that he or she must either work harder or cause the group to quit?
2. In that same vein, is the Köhler effect unique to persistence tasks, wherein the knowledge of one's indispensability comes precisely when one reaches one's ostensible performance limit, or does it also occur for a wider range of performance tasks?
3. In Köhler's (1926, 1927) study, group members cared about good task performance for personal, intrinsic reasons. In our paradigm, external incentives were provided. But is the effect moderated by the value group members place on group success?
4. In Köhler's study, group members probably identified strongly with the rowing club. In our paradigm, we tried to simulate this "social identity" by enhancing participants' connection to their gender group through a purported "battle of the sexes." But is such group identification necessary for, or facilitative of, the motivation gain effect?
5. Additionally, it seems likely that participants' subjective experiences play a role in how hard they try to accomplish the task. For instance, our present findings indicate that workers' affective reactions were related to motivation gains. A potentially fruitful issue for future work to pursue is the extent to which such affect is a basis for, versus a consequence of, task performance.

There are also a number of interesting external validity questions raised by our findings. One concerns the conditions under which the Köhler effect is and is not moderated by the discrepancy in group members' capabilities. At this point, we can only speculate about why Köhler observed an attenuation of motivation gains (see Figure 1) when discrepancy was large, whereas in our work we found no such effect. Although the two paradigms are (intentionally) quite similar, there are still a few differences. One such difference that might well be important is the participants' knowledge of one another's abilities. In Köhler's studies, the participants were members of a rowing club. They trained and competed together. Moreover, Köhler's performance assessments at his task were collected publicly at the club's facility (Köhler, 1926). These facts suggest that participants had a fairly complete and accurate knowledge of their relative abilities. In contrast, our participants were strangers when they arrived at the lab. Performance trials were conducted in cubicles with the experimenter and (in Experiment 1) a single other participant observing. For none of our dyad trials did a group member see both his/her own and his/her partner perform individually with the same hand prior to working in the dyad (see Table 2). Moreover, individual scores were not announced nor were observers or individual performers easily able to keep scores privately (recall that participants had to put their watches out of sight). All of these aspects of our paradigm suggest that our participants did not have a complete and accurate knowledge of their relative abilities. It seems plausible that one would be less likely to try to "keep up" with one's partner when one did know that one's partner was much more capable than when one didn't know precisely how capable one's partner really was. This conjecture could be simply tested in our paradigm by manipulating knowledge of partner ability prior to dyadic performance trials.

We also noted earlier that some studies (Kerr & MacCoun, 1984; Kerr & Sullaway, 1983) have observed distinctive motivation gains in mixed-sex groups. An open, but interesting, question is whether the sex composition of performance dyads would moderate the Köhler effect. For example, would a less capable male partner be willing to work especially hard not to be the first to quit when his partner was a woman rather than another man? Other group-composition variations would also be interesting to examine, for example, in- versus out-group partner, partners with versus without a continuing relationship. Yet another interesting question is how long the Köhler effect will persist. As we have noted, there is indirect evidence that group members see working hard and not failing at a conjunctive task as a challenge. Weaker members may enjoy demonstrating to themselves and to their partners that their true capabilities are higher than previously thought. However, once those true capabilities have been established, will weaker members continue to challenge themselves to meet them, or will they instead revert back to their old definition of "optimal" performance?

Of course, theory can be a powerful tool to guide inquiry. The I × V model suggests tentative answers to many of the open questions we have been noting. For example, it suggests that as

long as the weaker member sees his or her high performance as indispensable to the group (and values the personal and collective outcomes that follow from such high performance), motivation gains should result. This logic suggests that even though Köhler's and our paradigms—which both involve continuous performance feedback at a conjunctive persistence task—are particularly effective in producing such beliefs, it should be possible (in principle) to produce them with different group tasks as well.

And, although the I × V model shows promise for explaining the Köhler effect (as well as several other group motivation processes, cf. Karau & Williams, 1993, in press; Shepperd, 1993), there remain open theoretical questions as well. One is what valued outcomes drive the effect. If it is group success, per se, that is critical, then we would expect factors that affect the value placed on such success (e.g., intrinsic interest in the task; extrinsic rewards for group success; group identification) to powerfully moderate the effect. On the other hand, it may well be that the consequences of doing well at the conjunctive task for self-presentation are crucial. If so, factors that affect the value of positive self-presentation (e.g., present or future outcome interdependence; importance of the relationship; need for approval) should powerfully moderate the effect. Of course, it may be that both types of outcome contribute to the effect. One informative means of exploring this last question would be to vary group member anonymity. If group success, per se, is the crucial outcome, member anonymity should have no effect. If group members are more concerned with the impression they are making on fellow group members, we would expect a much stronger Köhler effect when members are personally identifiable.

In the last 25 years, since Ingham et al.'s (1974) seminal paper, social psychologists have shown that there are distinct and powerful motivation loss mechanisms that can undermine group performance. Köhler's (1926, 1927) seminal studies, along with more recent empirical and theoretical work (e.g., Karau & Williams, 1993, in press; Shepperd, 1993; Stroebe et al., 1996; Williams & Karau, 1991; the present studies), offer the hope that in the next 25 years our field will be equally successful in identifying powerful motivation gain mechanisms, with the promise of substantially improving the performance of work groups and teams.

References

Allport, G. W. (1924). *Social psychology.* Boston: Houghton Mifflin.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology, 59,* 177–181.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Baron, R. S. (1986). Distraction-conflict theory: Progress and problems. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 1–40). New York: Academic Press.

Baron, R. S., Kerr, N. L., & Miller, N. (1992). *Group process, group decision, group action.* Pacific Grove, CA: Brooks/Cole.

Erev, I., Bornstein, G., & Galili, R. (1993). Constructive intragroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology, 29,* 463–478.

Gaertner, S. L., Mann, J., Murrell, A., & Dovidio, J. F. (1989). Reducing intergroup bias: The benefits of recategorization. *Journal of Personality and Social Psychology, 57,* 239–249.

Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 8, pp. 47–99). New York: Academic Press.

Harkins, S., & Petty, R. E. (1982). Effects of task difficulty and task uniqueness on social loafing. *Journal of Personality and Social Psychology, 43,* 1214–1230.

Harkins, S., & Szymanski, K. (1989). Social loafing and group evaluation. *Journal of Personality and Social Psychology, 56,* 934–941.

Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Personality and Social Psychology, 10,* 371–384.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65,* 681–706.

Karau, S. J., & Williams, K. D. (in press). Understanding individual motivation in groups: The collective effort model. In M. E. Turner (Ed.), *Groups at work: Advances in theory and research.* Mahwah, NJ: Erlbaum.

Kerr, N. L. (1983). Motivation losses in task-performing groups: A social dilemma analysis. *Journal of Personality and Social Psychology, 45,* 819–828.

Kerr, N. L. (1990, October). *Reflections on group productivity.* Discussant in the Symposium "Group Performance" (W. Stroebe, Chair), Annual Convention of the Society of Experimental Social Psychology, Buffalo, NY.

Kerr, N. L., & Bruun, S. (1983). The dispensability of member effort and group motivation losses: Free rider effects. *Journal of Personality and Social Psychology, 44,* 78–94.

Kerr, N. L., & MacCoun, R. (1984). Sex composition of groups and member motivation II: Effects of relative member ability. *Basic and Applied Social Psychology, 5,* 255–271.

Kerr, N. L., & Sullaway, M. E. (1983). Group sex composition and member motivation. *Sex Roles, 9,* 403–417.

Kerr, N. L., & Yukelson, D. (1983). Group size and social facilitation. *Replications in Social Psychology, 3,* 6–9.

Köhler, O. (1926). Kraftleistungen bei Einzel- und Gruppenabeit [Physical performance in individual and group situations]. *Industrielle Psychotechnik, 3,* 274–282.

Köhler, O. (1927). Über den Gruppenwirkungsgrad der menschlichen Körperarbeit und die Bedingung optimaler Kollektivkraftreaktion [On group efficiency of physical labor and the conditions of optimal collective performance]. *Industrielle Psychotechnik, 4,* 209–226.

Kravitz, D. A., & Martin, B. (1986). Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology, 50,* 936–941.

Latané, B. (1981). The psychology of social impact. *American Psychologist, 36,* 343–356.

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,* 822–832.

Martin, L. L., Ward, D. W., Achee, J. W., & Wyer, R. S. (1993). Mood as input: People have to interpret the motivational implications of their moods. *Journal of Personality and Social Psychology, 64,* 317–326.

Matsui, T., Kakuyama, T., & Onglatco, M. L. U. (1987). Effects of goals and feedback on performance in groups. *Journal of Applied Psychology, 72,* 407–415.

Moreland, R. M., Hogg, M. A., & Hains, S. C. (1994). Back to the future: Social psychological research on groups. *Journal of Experimental Social Psychology, 30,* 527–555.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

Paulus, P. B., & Dzindolet, M. T. (1993). Social influence processes in group brainstorming. *Journal of Personality and Social Psychology, 64,* 575–586.

Ringelmann, M. (1913). Research on animate sources of power: The work

of man. *Annales de l'Institut National Agronomique, 2e serie-tome XII,* 1–40.

Ruess, M. (1992). *Ausdauerleistung in Dyaden: Eine Untersuchung zum Köhler-Effekt* [Persistence in dyads: A study of the Köhler-Effect]. Unpublished diploma thesis, University of Tübingen, Tübingen, Germany.

Schwarz, N., & Clore, G. (1996). Feelings and phenomenal experiences. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 433–465). New York: Guilford Press.

Shepperd, J. A. (1993). Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin, 113,* 67–81.

Steiner, I. D. (1972). *Group process and productivity.* New York: Academic Press.

Steiner, I. D. (1986). Groups and paradigms. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 251–289). New York: Academic Press.

Stroebe, W., Diehl, M., & Abakoumkin, G. (1996). Social compensation and the Köhler Effect: Toward a theoretical explanation of motivation gains in group productivity. In E. Witte & J. Davis (Eds.), *Understanding group behavior: Consensual action by small groups* (Vol. 2, pp. 37–65). Mahwah, NJ: Erlbaum.

Stroebe, W., Diehl, M., Abakoumkin, G., & Arnscheid, R. (1990, October). *The Köhler effect: Motivation gains in group performance.* Paper presented at the annual meeting of the Society of Experimental Social Psychology, Buffalo, NY.

Stroebe, W., & Frey, B. (1982). Self-interest and collective action: The economics and psychology of public goods. *British Journal of Social Psychology, 21,* 121–137.

Triplett, N. (1897). The dyanmogenic factors in pacemaking and competition. *American Journal of Psychology, 9,* 507–533.

Vroom, V. H. (1964). *Work and motivation.* New York: Wiley.

Williams, K. D., & Karau, S. J. (1991). Social loafing and social compensation: The effects of expectations of co-worker performance. *Journal of Personality and Social Psychology, 61,* 570–581.

Witte, E. H. (1989). Köhler rediscovered: The anti-Ringelmann effect. *European Journal of Social Psychology, 19,* 147–154.

Zajonc, R. (1965). Social facilitation. *Science, 149,* 269–274.

Appendix A

Ruess (1992) Simulations

A Monte Carlo similar to the one outlined in Appendix B was modeled on the design and procedure of Ruess (1992). In Ruess's study, there were 36 dyads, 18 tested with their dominant arms and 18 with their nondominant arms. In each session, dyad members that had performed together with one arm were also tested individually with the other arm. The mean individual performance scores Ruess observed were 276.6 s for the dominant arm and 253.7 s for the nondominant arm. It was not possible to estimate between performer and error variance directly, as in our simulation of Experiment 1. This was because there were not multiple individual performance scores collected for each participant. However, it was possible to do so indirectly by assuming that the ratio of between-subject to within-subject variance in individual performance in Ruess's study was identical to the ratio we observed in Experiment 1. Using standard ANOVA partitioning of variance, this ratio was found to be 6.63 (i.e., $SS_{\text{between persons}}/SS_{\text{within persons}} = 6.63$). Then, using the total variance in individual performance observed by Ruess, it was possible to likewise partition it into between- and within-subject components. This procedure resulted in estimates of $\sigma_{\text{persons}} = 60.96$ and $\sigma_{\text{error}} = 23.66$.

Due to the simpler design of Ruess's (1992) study, the corresponding simulation was also simpler than the simulation of Experiment 1 (see Appendix B). Using σ_{persons}, we first generated estimated true scores for a pair of participants (either for the dominant arm or nondominant arm). Then sampling from the standard normal distribution, as in Appendix B, two scores were generated for each dyad member—one individual trial score and one dyad trial score. From these, we could compute the ADS and CDS scores of interest. Ruess collected data from 36 dyads, so this procedure was repeated 36 times, with 18 dyads using their dominant arms and 18 using their nondominant arms. This constituted one experimental replication. The CDS and $r_{\text{ADS, CDS}}$ statistics were computed for each of 500 such experimental replications.

Again, the CDS statistic was unbiased; the mean CDS across experimental replications was .61, with the $\alpha = .05$ critical regions falling outside the interval $-9.98 \geq X_{\text{CDS}} \leq 11.13$. Thus, both by standard test procedures and the current simulation of the sampling distribution of CDS, the CDS of .78 observed by Ruess is clearly nonsignificant. More important for interpreting Ruess's findings, the mean $r_{\text{ADS, CDS}}$ across experimental replications was .265, with the $\alpha = .05$ critical regions falling outside the interval $-.0396 \leq r_{\text{ADS, CDS}} \leq .5444$. Thus, as reported in the text, the $r_{\text{ADS, CDS}}$ of .41 obtained for Ruess's sample was well within ($p = .35$) the range one would expect by chance alone.

Appendix B

Experiment 1 Simulations

The first Monte Carlo simulation was modeled on the design and procedure of Experiment 1. Thus, we assumed there were 6 participants in each experimental session, that each participant was tested individually with each arm, and that there were two subsets of 3 participants each and that within each subset, all (three) possible dyads were tested together with each arm (see Table 1). So, if we designate the 6 persons as A, B, C, D, E, and F, then we could take (A, B, C) and (D, E, F) as the two subsets. Thus, for each participant (e.g., A) we needed to generate two individual scores (e.g., A's individual scores with his/her dominant and his/her nondominant arm) and four dyadic scores (A's performance in the dyad [A, B] and the dyad [A, C], once with the dominant and once with the nondominant arm).

The following simple model was used to generate our scores:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad (1)$$

where μ = the grand mean for the particular (dominant or nondominant) arm, α_i = the departure of the i^{th} person's true mean from the grand mean, and ϵ_{ij} = an error component representing the departure of the j^{th} testing of the i^{th} person's performance from his/her true score ($\mu + \alpha_i$). Our best estimate of μ

was just the observed mean of individual performance (corrected for fatigue). In Experiment 1, for the dominant (dom) arm, $\mu = 174.8$ s, and, for the nondominant (nondom) arm, $\mu = 160.6$ s. We assumed that the distribution of individual true scores and error scores were distribed normally with mean of zero and standard deviations of σ_{persons} and σ_{error}, respectively. To estimate these standard deviations, we analyzed the two individual scores (corrected for fatigue) collected from each of the 64 participants in Experiment 1. Recall that our preliminary analysis of these scores indicated that there was no overall mean difference in performance with the dominant and nondominant arms. Thus, we might average these two scores to obtain our best estimate of each participant's true score and use the standard deviation of these estimates as σ_{persons}. Analyzing the data of Experiment 1, this procedure yielded an estimate of $\sigma_{\text{persons}} = 53.46$. Correspondingly, we might use the dispersion of the two individual scores (with the dominant and nondominant arm) around the true score estimate (their average) as an estimate of error variance. Pooling these deviation scores across the 64 participants of Experiment 1 led to an estimate of $\sigma_{\text{error}} = 20.69$.

With these estimates in hand, we could simulate the 12 individual (6 participants each tested with their dominant and nondominant arm) and 12 dyadic (three dyads in each of two subsets of 3 participants on both the dominant and nondominant arm) performances collected in each experimental session. Thus, for each arm of each person in each subset of participants, we would generate three separate scores—an estimated individual score and estimates of that person's performance in each of the two dyads in which he or she performed.

Let us illustrate for one subset of participants. For participant A, for example, in subset (A, B, C) using the dominant arm, we first estimated A's true score, $Y_{\text{A-dom}}$, by sampling a value from the standard normal distribution ($M = 0$, $SD = 1$) using the Random Number Generation tool in the Excel spreadsheet program; let us call that value $\tau_{\text{A-dom}}$.[A1] Applying the assumed model of estimation,

$$Y_{\text{A-dom}} = \mu + \alpha_i = \mu + \tau_{\text{A-dom}}\sigma_{\text{persons}} = 174.8 + 53.46\tau_{\text{A-dom}}. \quad (2)$$

Then, three more values were independently sampled from the standard normal distribution, $\tau_{\text{A-dom,individual}}$, $\tau_{\text{A-dom,A-B dyad}}$, and $\tau_{\text{A-dom,A-C dyad}}$.[A2] Next, the simulated performances of participant A with his or her dominant arm were simply defined as

$$Y_{\text{A-dom,individual}} = Y_{\text{A-dom}} + \epsilon_{\text{A-dom,individual}} = Y_{\text{A-dom}} + \tau_{\text{A-dom,individual}}\sigma_{\text{err}}, \quad (3)$$

$$Y_{\text{A-dom,A-B dyad}} = Y_{\text{A-dom}} + \epsilon_{\text{A-dom,A-B dyad}} = Y_{\text{A-dom}} + \tau_{\text{A-dom,A-B dyad}}\sigma_{\text{err}}, \quad (4)$$

$$Y_{\text{A-dom,A-C dyad}} = Y_{\text{A-dom}} + \epsilon_{\text{A-dom,A-C dyad}} = Y_{\text{A-dom}} + \tau_{\text{A-dom,A-C dyad}}\sigma_{\text{err}}. \quad (5)$$

The same procedure was followed for the other two participants with whom A was paired in a dyad, B and C (i.e., we simulated [$Y_{\text{B-dom,individual}}$, $Y_{\text{B-dom,A-B dyad}}$, and $Y_{\text{B-dom,B-C dyad}}$) and ($Y_{\text{C-dom,individual}}$, $Y_{\text{C-dom,A-C dyad}}$, and $Y_{\text{C-dom,B-C dyad}}$]). Because the overall mean for the nondominant arm was 14.2 s less than for the dominant arm, we assumed that the true score for the nondominant arm was just 14.2 s less than the estimated true score for the dominant arm; that is, $Y_{\text{i-nondom}} = Y_{\text{i-dom}} - 14.2$. We then used the same procedure to simulate individual and group performance scores (i.e., $Y_{\text{A-nondom,individual}}$, $Y_{\text{A-nondom,A-B dyad}}$, $Y_{\text{B-nondom,E-F dyad}}$) for the nondominant arm.

For each dyad, we then computed the discrepancy of member abilities, ADS, and the discrepancy between the dyadic performance and the weaker individual performance of the two dyad members, CDS. For example, for the dyad A–B using their dominant arm,

$$\text{ADS}_{\text{A-B dom}} = \max(Y_{\text{A-dom,individual}}, Y_{\text{B-dom,individual}}) - \min(Y_{\text{A-dom,individual}}, Y_{\text{B-dom,individual}}) \quad (6)$$

$$\text{CDS}_{\text{A-B dom}} = \min(Y_{\text{A-dom,A-B dyad}}, Y_{\text{B-dom,A-B dyad}}) - \min(Y_{\text{A-dom,individual}}, Y_{\text{B-dom,individual}}) \quad (7)$$

In all, 12 such pairs were calculated in each session, one pair per dyad (A–B dom, A–C dom, B–C dom, D–E dom, D–F dom, E–F dom, A–B nondom, A–C nondom, B–C nondom, D–E nondom, D–F nondom, E–F nondom)

In Experiment 1 there were 14 sessions, so this same basic procedure was replicated 14 times. This iteration resulted in 168 pairs of (ADS, CDS) scores, corresponding exactly to the sample size of Experiment 1. Hence, this procedure constituted one experimental replication. Again, the statistics that we were most interested in were the mean CDS score and the correlation between ADS and CDS scores, $r_{\text{ADS, CDS}}$. These two statistics were computed for the experimental replication. To identify means and critical values of the sampling distribution of CDS and $r_{\text{ADS, CDS}}$, 500 experimental replications were conducted.

As asserted in the text, our simulation confirmed that CDS was an unbiased index of motivation gain. The mean of CDS estimates across the 500 replications was very nearly zero (−.075); the $\alpha = .05$ critical regions fell outside the interval $-4.64 \geq X_{\text{CDC}} \leq 4.56$. (The latter was simply determined by locating the 2.5 and 97.5 percentiles of the distribution of 500 CDS scores.) Thus, both by standard null hypothesis testing and by our simulation results, the CDS of 14.25 is significantly larger than one would expect by chance.

We also asserted in the text that regression artifacts would bias $r_{\text{ADS, CDS}}$; that is, even when group members performed no differently in dyads than when performing individually, as long as there was some random error in performance across performance occasions, dyads with relatively larger discrepancies in observed individual performance would also tend to show relatively more motivation gain (relative to its less capable member). And indeed, the mean of the $r_{\text{ADS, CDS}}$ estimates across the 500 replications was .250; the $\alpha = .05$ critical regions fell outside the interval $.088 \leq r_{\text{ADS, CDS}} \leq .406$. Thus, the $r_{\text{ADS, CDS}} = .19$ observed in Experiment 1 can be easily attributed to the effects of chance.

It was also asserted in the text that the index of motivation gain utilized by Köhler, the ABR = 100 × dyad/(average individual score), was biased—that is, that regression artifacts tended to produce ABR scores that are smaller than the nominal 100 no-motivation-gain baseline. To confirm this, we also computed ABR. As expected, the mean ABR score was 81.3, and the values expected by chance are well below the nominal no-motivation-gain baseline of 100 (the $\alpha = .05$ critical regions fell outside the interval $76.1 \leq \text{ABR} \leq 86.0$).

[A1] Actually, there was one additional assumption in our model—that the distribution of member abilities and error scores was not unbounded (as in the normal distribution), but was bounded. It did not make sense, for example, for our estimate of a person's true score to be less than zero. However, if no constraints were placed on sampling from the unit normal distribution, such impossible scores could occur. Likewise, scores too high to actually be obtained by real human participants could result from sampling from the unbounded unit normal. Inspecting our individual performance scores in Experiment 1, it was determined that the highest and lowest scores fell (depending on arm) between two and three standard deviations above and below the observed mean. We therefore imposed bounds on our sampling of the unit normal, such that sampled τ values above 2.5 were set equal to 2.5, and sampled τ values below −2.5 were set equal to −2.5.

[A2] For the same reasons outlined in Footnote 12, if $\tau < -2.5$, τ was set equal to −2.5, and if $\tau > 2.5$, it was set equal to 2.5.

Received June 11, 1999
Revision received April 12, 2000
Accepted April 18, 2000 ■