Journal of Experimental Psychology: General

© 2020 American Psychological Association
ISSN: 0096-3445

2020, Vol. 150, No. 5, 890–914
http://dx.doi.org/10.1037/xge0000975

A Test of Synergy in Dynamic System Control Tasks

Thomas Schultze
University of Göttingen and Leibniz ScienceCampus Primate Cognition, Göttingen, Germany

Sylvana Drewes
University of Göttingen

Stefan Schulz-Hardt
University of Göttingen and Leibniz ScienceCampus Primate Cognition, Göttingen, Germany

Individual performance in controlling complex dynamic systems such as managing production in a company or keeping ecosystems in balance is often suboptimal. In this article, we provide the first unequivocal test of whether groups are superior to individuals when controlling dynamic systems. In addition, we test to what extent performance advantages of groups are simply the result of statistically aggregating a larger number of individual opinions and to what extent they represent true synergy attributable to within-group interaction. In 3 experiments, we compared the system control performance of interacting real groups with that of equally sized nominal groups and with individuals. We provide evidence that groups indeed perform better than individuals in dynamic system control tasks. Furthermore, in comparing real groups with nominal groups, we show that, although the majority of real groups' performance advantage stems from statistical aggregation, there is also evidence of true synergy. Finally, we identify the mechanism by which groups achieve synergy, namely group-to-individual transfer. Discussion allows group members to exchange critical information about the system, leading to an improved individual capability to control the system, which, in turn, improves group performance.

Keywords: complex problems, dynamic systems, group decision-making, group performance, synergy

Supplemental materials: http://dx.doi.org/10.1037/xge0000975.supp

Dynamic systems are sets of variables that interact in a time-dependent fashion; that is, the current state of the system depends on previous system states. Dynamic system control tasks, sometimes also referred to as complex problem solving, pertain to numerous domains of private and professional life. Some control tasks, like maintaining a healthy body weight by adjusting food choices and the amount of exercise, are private in nature. Other tasks, such as stabilizing the cooling system of nuclear power plants in the face of disturbances or managing companies, are important to ensure public safety or economic success. Inadequate system control decisions have contributed to catastrophes like the explosion of the Chernobyl reactor (cf. Hofinger, Rek, & Strohschneider, 2006; Reason, 1987), and mismanagement of dynamic business systems fosters severe monetary losses that might lead to business failures and job losses (Fredrickson & Mitchell, 1984).

One popular remedy to ensure high-quality decisions is the employment of groups. Consistent with this notion, groups are becoming the most frequent decision-making entity in many organizations (Devine, Clayton, Philips, Dunford, & Melner, 1999), particularly in the context of managing dynamic systems (Kluge, 2014). On a general level, groups can outperform individuals for two reasons. The first is a purely quantitative effect: Groups have superior numbers. Just as we would expect a group of three people to be able to lift heavier objects than an individual because, collectively, they possess greater physical strength, we would expect groups to have an advantage over individuals when working on cognitive tasks because, combined, they possess greater cognitive and attentional resources as well as greater and more diverse knowledge. These greater resources should lead to better performance even in the absence of any group interaction. In the context of our study, and in line with previous research, we will describe performance advantages of groups that are merely the result of their superior numbers as *statistical aggregation effect* (Zajonc, 1965). Second, groups can achieve performance gains beyond the statistical aggregation effect as the result of the group interaction, for example by integrating their members' individual knowledge to gain novel insights, or because of social learning within the group. Such performance advantages attributable to group processes denote qualitative differences between individual

This article was published Online First October 12, 2020.

Thomas Schultze, Institute of Psychology, University of Göttingen, and Leibniz ScienceCampus Primate Cognition, Göttingen, Germany; Sylvana Drewes, Institute of Psychology, University of Göttingen; Stefan Schulz-Hardt, Institute of Psychology, University of Göttingen, and Leibniz ScienceCampus Primate Cognition.

The data and analyses scripts necessary to reproduce our results as well as the materials are publicly available at the Open Science Framework (https://osf.io/e3uky).

The data reported here were presented at the 14th INGRoup conference in Lisbon, Portugal, in July 2019.

Correspondence concerning this article should be addressed to Thomas Schultze, Institute of Psychology, University of Göttingen, Goßlerstraße 14, D-37073 Göttingen, Germany. E-mail: schultze@psych.uni-goettingen.de

and group work. In social psychological group performance research, these benefits of group interaction are called process gains (Hackman & Morris, 1975). However, because the relevance of dynamic system control transcends social psychology, we will use the more common term *synergy* here. Note, that the group process can also be detrimental for a group's performance (e.g., Ingham, Levinger, Graves, & Peckham, 1974; Latané, Williams, & Harkins, 1979; Ziegler, Diehl, & Zijlstra, 2000). Thus, group performance might fall short of what we would expect based on the statistical aggregation effect, and, in the worst case, not differ from the average individual performance. We will term groups falling short of the expected performance under the statistical aggregation effect as *antisynergy*.

In line with the saying that "two heads are better than one," there is already evidence of both statistical aggregation effects and true synergy in simple judgment and decision tasks, such as choices between different options (for a review, see Schulz-Hardt & Mojzisch, 2012), quantitative judgments (Minson, Mueller, & Larrick, 2018; Schultze, Mojzisch, & Schulz-Hardt, 2012; Sniezek & Henry, 1989; Stern, Schultze, & Schulz-Hardt, 2017), and logical problems (Laughlin, Hatch, Silver, & Boh, 2006; Laughlin, Zander, Knievel, & Tan, 2003). However, comparable evidence is lacking in the context of the more complex dynamic systems. Given the high relevance of successful dynamic system control performance for public safety and prosperity, the pertinent question of whether we should rely on groups for system control needs clarification. Therefore, we aim to close this research gap by empirically investigating the possible benefits (or disadvantages) of group work for control performance and by providing an unequivocal test of synergy in dynamic system control.

In the following, we outline what defines dynamic systems, how dynamic system control is studied in the laboratory, and what makes dynamic systems difficult to handle. We then turn to groups as the decision-making units. There, we first briefly review the existing studies comparing group and individual performance and show why these studies do not yet provide an unequivocal test of our research question. Next, we describe the theoretical framework we use to disentangle statistical aggregation effects and synergy in group decision-making, and we theorize on the specific types of group processes that might lead to synergy in dynamic system control tasks. Using this framework, we develop a research design that allows us to both measure the performance advantage of groups over individuals and decompose this performance advantage into effects of statistical aggregation, on the one hand, and true synergy, on the other. Using this newly developed research design, we then test for group superiority and, potentially, synergy in dynamic system control in a series of three experiments.

Dynamic Systems

According to Edwards (1962), dynamic systems control tasks have three characteristics: First, a series of decisions are required to reach the goal of achieving and maintaining a desired system state. Second, these decisions are time-dependent, meaning that later decisions are constrained by earlier ones. Finally, the system has its own dynamics, that is, the state of the system changes not only in response to the decision maker's interventions, but also does so autonomously. To make matters worse, the relations between the system's variables are often intransparent, nonlinear, and time-delayed (Funke, 1991). In addition, there may be hidden variables that link the control variables to the targets but are, themselves, unobservable for the person controlling the system. However, observing changes in the system states allows the acquisition of system knowledge by testing hypotheses about the relation between certain system variables. That is, given sufficient experience, the person controlling the system can figure out whether and how the control and target variables are connected.

Performance in dynamic system control is usually investigated using computer simulations of technical, biological, or economic systems (these simulations are sometimes called microworlds). Participants either have the objective to reach a certain equilibrium state such as the optimal temperature in a cooling system or to maximize a target variable such as profit by minimizing costs in a business system. To accomplish these goals, participants make repeated decisions about adjusting one or more control variables. For example, in a business simulation, participants may decide to change production rates in a factory to satisfy changing demands. After each decision, they receive information about the current system state such as the sales, inventory, and current resulting costs of the company they manage. The most common way to assess system control performance is to compare the current system state with an optimal benchmark; that is, the best possible approximation of the respective goal (attaining equilibrium or minimizing/maximizing an outcome).

Early research on dynamic system control performance used simulations that aimed for mundane realism, which often translated to complex dynamic systems with many interconnected system variables. A prominent example is the business simulation TAILORSHOP with 24 system variables, in which participants' task is to maximize profit by buying resources and modifying production capacities (Putz-Osterloh, 1981). Another simulation frequently used in early research on dynamic system control is the foreign aid scenario TANALAND (53 system variables), in which participants can take several measures to improve the life of the inhabitants of a fictional African country (Dörner & Reither, 1978). The pinnacle of complexity is certainly the LOHHAUSEN simulation, in which participants took the role of the mayor of a small German town and which consists of roughly 2,000 system variables (Dörner, Kreuzig, Reiter, & Stäudel, 1983). However, some early studies on dynamic systems also used smaller systems such as COLDSTORE (Reichert & Dörner, 1988), a scenario with only six system variables, in which participants need to stabilize the room temperature in a retail store to prevent their dairy products from either freezing or becoming too warm. Common features of these early system control tasks are nonlinear and time-delayed variable relations, side effects of the system control decisions, own dynamics, and hidden variables. We refer the interested reader to Funke (1991), who provides an exhaustive overview of the system control tasks used in early research on dynamic system control performance.

More recently, research on dynamic system control sacrificed some of the mundane realism—and thus complexity—of earlier systems for the sake of better psychometric characteristics (Dörner & Funke, 2017). On average, system control tasks used nowadays tend to have fewer system variables, which also made using them more economically feasible (consider that participants who worked on the LOHHAUSEN simulation spent around 16 hr in the lab). In addition, variable relations in more recent dynamic system control tasks are frequently linear and transparent, and they often lack time-delays and side effects. For example, Osman and colleagues used a dynamic system in which participants controlled three input variables in order to optimize one outcome variable (Osman, Glass, & Hola, 2015; Osman, Glass, Hola, & Stollewerk, 2017). In this system, the outcome is a linear combination of the prior value of the outcome variable, two of the three input variables, and added noise, which represented the system's own dynamics. Another example is the MicroDYN toolset, which allows creating relatively simple (or, in terms of the authors, minimally complex) dynamic systems. In one of its first applications, Greiff, Wüstenberg, and Funke (2012) created a set of dynamic systems for participants to work on. In each system, there was a set of four input and four output variables. Each output variable was a linear combination of its prior state (and, in a few cases, the prior state of a second outcome variable) as well as one or two of the input variables with no added noise.

It is perhaps not surprising that individual performance in dynamic system control varies with the respective task. Common findings in earlier studies were that a—usually small—group of participants performed well while the others managed to destabilize the system, with the result of catastrophic failure (Brehmer, 2005; Sterman, 1989a; for an overview see also Rouwette, Größler, & Vennix, 2004). For example, a typical outcome of TANALAND was the near-extinction of the simulated population (Dörner & Reither, 1978), and in COLDSTORE, participants often created highly volatile systems, in which the dairy products were repeatedly overheated and frozen (Reichert & Dörner, 1988). Those early studies created the impression that individual decision-makers perform poorly when dealing with dynamic systems. However, this conclusion may not be generally true. Some of the more recent studies paint a more optimistic picture of system control performance, suggesting that individuals can manage less complex dynamic systems quite well, even with rather impoverished information (e.g., Osman et al., 2015, 2017).

Given the variability in individual system control performance, the question is what makes (some) dynamic system control tasks difficult. Adopting the conceptualization by Osman and Palencia (2019), we can generally describe dynamic system control tasks as dynamic decision-making under uncertainty, with the degree of uncertainty about the input-output relations determining task difficulty. This uncertainty increases with the number of system variables and their interconnections (Funke, 1985, 1992). It also increases as feedback becomes more time-delayed and indirect (Funke, 2012) and to the extent that the system has its own dynamics and side effects (Paich & Sterman, 1993; Sterman, 1989b).

In line with the conceptualization of dynamic system control as decision-making under uncertainty, the key to managing dynamic systems is to reduce this uncertainty by acquiring and applying causal knowledge about the system (Osman, 2010). To this end, decision-makers can engage in monitoring and controlling behaviors. Monitoring means that decision-makers track changes in the system and develop testable hypotheses about the variable relations, but it also entails tracking whether the current strategies of the decision-maker in terms of system inputs produce the desired outcomes. Controlling, in contrast, refers to goal-directed input decisions. Assuming that these decisions reflect the decision-maker's current system knowledge, they allow testing hypotheses about variable relations and generate feedback that allows updating the system knowledge. As Osman lines out, decision-makers working on dynamic systems face an exploration-exploitation problem, in which they first need to explore the system to understand its causal structure and can, then, exploit their system knowledge when aiming to produce the desired system states. Early studies on complex dynamic system control, in which the behavior of individual participants was evaluated qualitatively, support this idea. Specifically, it seems that participants perform particularly well if they explore the system carefully initially and focus more on monitoring than on controlling (e.g., Dörner et al., 1983; Reichert & Dörner, 1988). In contrast, poor performance when controlling dynamic systems can usually be traced back to systematic misperceptions of system characteristics, that is, insufficient system knowledge (Funke, 1992). Having established that the key to successful dynamic system control is the acquisition of system knowledge, we now turn to the question whether groups are superior to individuals at managing dynamic systems.

Group Versus Individual Performance in Dynamic System Control

Given that individual decision-makers often perform poorly when the system they work on exceeds a certain level of complexity, one prominent idea is to turn to groups as the decision-making unit. Unfortunately, empirical research investigating whether groups are superior to individuals when working on dynamic systems is scarce, and the results are mixed. Here, we briefly review the existing evidence and elaborate why it does not yet allow drawing firm conclusions about the performance of groups relative to individuals when controlling dynamic systems.

Essentially, we can identify two clusters of studies comparing group and individual system control performance. The first is relatively pessimistic regarding groups' ability to outperform individuals. In the earliest of these studies, Dentler (1977) asked participants to work on the above-mentioned foreign aid scenario TANALAND, treating the number of input decisions and system stability as performance measures. Endres and Putz-Osterloh (1994) used a variation of the same system to compare the performance of three-person groups to that of individuals. Badke-Schaub (1993) also compared the performance of three-person groups with that of individuals but used a simulation in which participants' goal was to prevent new AIDS infections in a major city. In order to succeed in this simulation, participants had to consider delayed effects of newly introduced measures due to the incubation period as well as the exponential rate at which the disease spreads. Two other studies investigated group versus individual system control performance using variants of the TAILORSHOP business simulation. Köller, Dauenheimer, and Strauß (1993) had participants manage the textile factory either as individuals or as dyads, while Leutner (1988) compared the performance of three-person groups with that of individuals. Neither of these studies found significant differences between individuals and groups in terms of system control performance.

The second set of studies draws a more positive picture of groups' dynamic systems control performance. In one of these studies, participants managed a water purification plant either individually or as dyads (Gonzalez, Thomas, & Vanyukov, 2005). Their goal was to deliver the correct amount of purified water to different clients at a given time. The core challenges of this scenario were resource limitations and nonlinear variable relations. Also comparing dyads with individuals, Kirlik, Plamondon, Lytton, Jagacinski, and Miller (1993) used a command and control scenario in which participants had to steer a helicopter on a map containing both friendly and hostile targets as well as cargo that needed to be hauled to the home base. The aim was to destroy hostile targets (while avoiding destroying friendly targets) and to bring home as much cargo as possible using a limited amount of fuel and ammunition. This task focused on efficient resource management and planning complex routes due to moving targets, but it did not entail time delays or nonlinear variable relations. Finally, Wolfe and Chacko (1983) compared individuals with dyads, three-person groups, and four-person groups in a sophisticated business simulation. In this simulation, participants managed a company and could make multiple decisions on each turn concerning, for example, products, prices, production, sales, acquisitions, and plant expansions with the goal of maximizing profit. In all three studies, groups outperformed individuals significantly.

In sum, the two sets of studies imply that groups either perform at the level of individuals or that they outperform them. However, all of these prior studies suffer from a methodological shortcoming that precludes clear-cut conclusions: The decision-making entity (group vs. individual) was always confounded with, at least, one other factor that might have been highly relevant to the performance in the control task. Specifically, there were confounds of the type of decision maker with the amount of information available (Wolfe & Chacko, 1983), the number of control decisions to be made (Dentler, 1977; Gonzalez et al., 2005), the number of control devices (Kirlik et al., 1993), task experience (Endres & Putz-Osterloh, 1994), goal setting (Badke-Schaub, 1993; Leutner, 1988), and even the decision task itself (Köller et al., 1993).

Therefore, it remains to be seen whether group performance surpasses individual performance in a fair comparison. As such, one of the aims of our study was to provide a fair test devoid of confounding variables. However, even in the absence of confounds, comparing groups with individuals is not entirely informative, because it does not allow distinguishing between a statistical aggregation effect, on the one hand, and true synergy (or lack thereof), on the other. This distinction is not only relevant for theory building by shedding light on the synergetic potential of real group discussion, but also for practical purposes. If, for example, groups were superior to individuals solely because of their superior numbers, and not at all because of group processes, it may be more efficient to rely on statistical aggregation of individual control decisions instead. However, disentangling these effects requires a specific research design, which we will describe in the following.

Separating Synergy From the Statistical Aggregation Effect

As mentioned initially, groups might outperform individuals as decision-makers in dynamic system control because of a quantitative effect (groups encompass multiple people), or a qualitative effect (group interaction leads to synergy), or a combination of the two. To elaborate, we draw on a general framework of group performance (Hackman & Morris, 1975). The underlying idea of this framework is that group performance is the sum of two components: the individual component, which depends on the number of group members as well as their individual motivation and ability to perform the task, and the group component, which reflects the sum of all group processes that influence the group's performance.

The individual component of group performance is what Steiner (1972) termed the *group potential*. It represents the group performance we would expect absent any positive or negative effects of social interaction within the group. Put differently, the group potential is a theoretical performance benchmark that assumes that the group component is zero. The most common way to calculate the group potential is to use so-called *nominal groups*. Nominal groups consist of a number of individuals that work independently and whose individual performance is later combined in a sensible way to create the nominal group performance. Thus, nominal group decisions are noncollaborative decisions while real group decisions are collaborative.[1] The proper way to compute the group potential from nominal group members' individual performance depends on the respective group task. Here, we are concerned with quantitative decision tasks, in which groups must decide on a specific level of a given quantity on a continuous scale rather than choosing between distinct categories. In the case of such quantitative decision tasks, the group potential is the average of the group members' individual decisions (Steiner, 1966). Note here, that the choice of the mean of nominal group members' individual contributions as the group potential is theory-driven. It reflects the most plausible strategy to combine quantitative decisions absent any interaction within groups. Other aggregation strategies such as weighting nominal group members' individual contributions by relative accuracy (Bonner, Sillito, & Baumann, 2007) or ignoring the weaker nominal group members (Mannes, Soll, & Larrick, 2014) might yield better performance and, thus, an even more challenging performance benchmark (see also Davis-Stober, Budescu, Dana, & Broomell, 2014). However, these aggregation strategies do not serve well when determining the purely individual component of group performance, because they require knowledge about relative expertise that groups can, realistically, only obtain via interaction.

Naturally, the group potential increases with the group members' individual ability and motivation to perform the task, that is, ceteris paribus, we would expect groups comprising skilled and motivated individuals to outperform groups with less able and less motivated members. More importantly, the group potential in decision-making also increases with the number of group members due to the aforementioned statistical aggregation effect. Essentially, each group member's individual decision can be decomposed into the true value, an idiosyncratic bias, and random error (Yaniv, 2004). In the context of dynamic system control, such idiosyncratic biases reflect systematic misconceptions about the causal relations of the system variables, whereas the errors reflect imprecision that may stem from both incomplete system knowledge but also from a system being noisy. Averaging group mem-

[1] For the sake of completeness, we should mention that, as long as group members have not yet interacted with each other, it is also possible to compute the group potential based on their individual prediscussion performance by treating future group members as if they were a nominal group. However, because this approach is not applicable in our study, we do not consider it further.

bers' individual decisions allows cancellation of both their idiosyncratic biases and errors, an effect also known as the wisdom of the crowds (Surowiecki, 2004).[2] The benefit of such aggregation effects increases with group size, and because we can conceive of individuals as one-person groups, we would also expect the aggregate decision of multiple group members to be more accurate than the average individual decision (for a formal analysis, see Einhorn, Hogarth, & Klempner, 1977).

Having established that groups can plausibly outperform individuals in dynamic system control tasks by superiority of numbers alone, we now turn to the group component of group performance. We can compute this component by subtracting the group potential (i.e., the performance of noninteracting nominal groups) from interacting groups' actual performance. If the result is positive, then social interaction added something to the groups' performance beyond an aggregation effect, that is, the group achieved synergy. As mentioned initially, it is also conceivable that group processes that are detrimental to the group's performance outweigh or even override the synergetic processes, leading to antisynergy.

Because dynamic system control consists of a sequence of decisions in the same context, we approach the search for synergy from the perspective of group learning, that is, group-specific learning processes that exceed simple individual training effects. According to the dynamic model of group performance by Brodbeck and Greitemeyer (2000), we can distinguish two general categories of group learning processes. First, performing a task collaboratively in a group context can enhance one's individual task-related competencies, resulting in individual capability gains. For example, the group member with the most accurate knowledge about the system's causal structure can convey this knowledge to the other group members. Alternatively, interacting groups might be better at monitoring a dynamic system because, collectively, they can focus their attention on more system variables, and they can discuss their observations, allowing them to correct each other's misconceptions. Such processes are usually labeled as *group-to-individual transfer* (*G-I transfer*; see, e.g., Laughlin, Carey, & Kerr, 2008). G-I transfer improves the group members' individual knowledge and manifests as improved system control performance of real group members relative to individuals or members of nominal groups, that is, their idiosyncratic biases and errors become smaller. Accordingly, G-I transfer improves the performance of real groups relative to that of nominal groups and individuals, because the group decision is now some combination of already more accurate individual contributions.

Second, repeatedly performing similar tasks in a group can also improve the group's ability to coordinate. For example, group members might develop metaknowledge about particular skills, strengths, and weaknesses of the other group members, so that one knows whom to ask if something is unclear (a *transactive memory system* in the terms of Wegner, 1986). Learning about the relative expertise of group members allows for better coordination via *differential weighting* of group members' inputs (Stern et al., 2017). Consider a case in which members of a three-person group have learned that one of their members consistently makes better predictions about how the system state is going to change as a function of the joint input decision. The group might then weight this member's individual input more—or even adopt it as the group decision—in subsequent rounds of the system control task (weighting by expertise). Alternatively, mutual error checking might allow groups to identify individual contributions that are particularly bad given the group's current system knowledge, irrespective of which member suggested it. The groups can then assign less weight to such inaccurate individual contributions or even neglect them when making a joint decision (weighting by accuracy). Weighting group members' individual contributions by expertise or accuracy should lead to groups (further) outperforming nominal groups and individuals.

Based on this reasoning, we can now evaluate the system control performance of groups properly. If real groups outperform individuals working on dynamic system control tasks, we first need to compare nominal groups with individuals. If nominal groups outperform individuals, we can attribute their superiority to the statistical aggregation effect. Next, we need to compare the performance of interacting real groups with that of nominal groups. If the performance of interacting groups surpasses that of nominal groups, this difference indicates true synergy because it can only result from group processes.

A Nominal Group Design for Dynamic System Control Tasks

Whereas comparing the performance of groups with that of individuals is as straightforward in dynamic systems as it is in simpler quantitative decision tasks, computing the performance of a nominal group is more challenging. In nondynamic tasks (i.e., all decisions are independent of each other), building a nominal group condition as a comparison for the interacting real group condition would simply mean to have an equivalent number of participants in an experimental session who make the quantitative decisions independently, then average these decisions, and finally compute the performance resulting from this aggregate (e.g., Schultze et al., 2012). Nominal groups would then benefit from statistical aggregation, but not from any group processes. However, in a dynamic system control task, each decision to alter one or more of the input variables changes the state of the system for all trials to come. For example, if one person decided in the first trial to increase a certain input variable while another person decided to decrease it, both would end up with different states of the system right after the first trial. Therefore, from the second trial on, the two individuals would control different systems, that is, the exact same decisions would now produce different results or, put differently, achieving the desired system state would require different decisions. Statistical aggregation of several individuals who control different systems amounts to the proverbial comparison of apples and oranges. Thus, the challenge in creating a nominal group condition in dynamic system control is to ensure that all nominal group members face the same system on each trial, while at the same time preventing nominal group members from communicating or interacting with each other.

Our solution to this problem is a nominal group condition where, in each round, the nominal group members make individual

[2] Because we focus on quantitative decisions here, we do not consider categorical decisions (e.g., decisions between options A, B, and C). However, for those types of decisions the same logic applies, with the only difference that aggregation takes the form of majority voting instead of averaging.

decisions about the input variables. These judgments are then averaged (by a server computer to that the three participants' terminals are connected—for details, see the methods section), and this average serves as the system input for this round. Afterward, this aggregate input as well as the resulting change in the system's state are reported to the nominal group members, and the next round begins. Omitting feedback about the aggregated control decision is not an option, because this would imply that nominal group members falsely attribute changes in system states to their individual control decision rather than the nominal groups' aggregate decision, thus leading to erroneous system knowledge and bad performance. We are aware that, with this procedure, participants in this nominal group condition still experience some sort of rudimentary interaction. Being informed about how the input variable was altered inevitably gives them some idea about what control decisions the other group members had made, even though they do not learn the exact values the other nominal group members suggested. Because there is some evidence that mere awareness of others' decisions can induce a certain amount of G-I transfer (Farrell, 2011; Maciejovsky & Budescu, 2007), this means that even nominal group members in our design may benefit from synergy to some extent (which makes it somewhat more difficult to detect synergy in the real group condition). However, this condition is the closest to a pure nominal group condition that we could think of, given the constraints inherent to dynamic system control tasks.

The Present Research

The present research consisted of three experiments and focused on three major goals. First, we aimed at providing an unequivocal comparison of group versus individual performance in dynamic system control tasks. Therefore, we compared the control performance of interacting three-person-groups with that of individuals. Second, given evidence of the superiority of groups, we aimed to test for true synergy beyond the statistical aggregation effect by comparing the performance of interacting real groups with that of noninteracting nominal groups. Finally, given evidence of synergy, we aimed to investigate to what extent it stemmed from G-I transfer, on the one hand, and/or differential weighting, on the other. We tested for possible benefits of differential weighting in all three experiments by comparing real groups' actual performance to the hypothetical performance that would have resulted from averaging the real group members' individual prediscussion decisions for each round. Real group decisions producing better outcomes than merely averaging their members' contributions constitutes evidence of synergy by differential weighting. Experiment 3 also included a postgroup phase allowing us to test for G-I transfer by comparing the individual performance of participants who had worked individually before with that of former nominal and real group members. We can conclude that there is synergy due to G-I transfer if former real group members perform better individually in the postgroup phase than former nominal group members and individuals.

Method

Because of large similarities in sample characteristics, design, and methodological approach of the three experiments in our study, we present a common methods section for all three experiments. Descriptions of common characteristics in each method section are supplemented by special features of the single experiments. The complete written instructions for all three experiments in both German (original version) and English (translation) are available online at the Open Science Framework at https://osf.io/e3uky (Schultze & Schulz-Hardt, 2020).

Participants and Design

In each of the three experiments, 315 undergraduate and graduate students of a German university took part. The sample size, which corresponds to 35 real and nominal groups and 35 triples of individuals per experiment, was not based on a power analysis but rather reflects the upper end of the range of sample sizes used in previous group performance research (e.g., Laughlin et al., 2008). Participants were recruited on campus and via the participant database ORSEE (Greiner, 2004). In Experiment 1, 179 of the participants were female (56.80%), and participants' average age was 23.04 years (*SD* = 3.13 years). Experiment 2 comprised 187 female participants (59.40%), and participants' average age was 22.89 years (*SD* = 2.76 years). In Experiment 3, 172 of the participants were female (54.60%), and the average age of participants was 23.52 (*SD* = 3.75 years). Participation was voluntary, and we informed participants at the beginning of each experimental session that they could withdraw from the experiment at any time without fear of consequences. We also informed them that data would be collected and analyzed in an anonymized fashion. Additionally, participants in Experiment 3 provided written consent for video recordings of the real group discussions.

Experiments 1 and 2 used a 2 × 3 design with the within-subject factor *test phase* (Phase 1 vs. Phase 2) and the between-subject factor *decision-maker* (individual vs. nominal group vs. real group). In Experiment 3, we added a third phase to test for the occurrence of G-I transfer. In each of the experiments, the between-subjects factor was manipulated in Phase 2.

Experimental Tasks and Measures

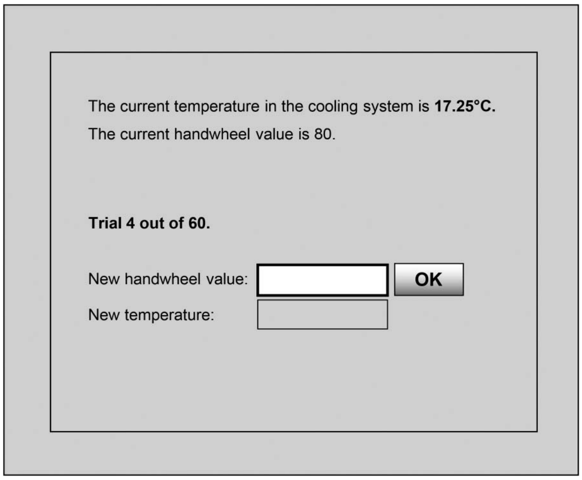
Experiments 1 and 3. The dynamic control task of Experiments 1 and 3 was an adapted version of the COLDSTORE simulation mentioned above (Reichert & Dörner, 1988). In the original study, participants were to control the cooling system of a cold storage house containing dairy products. We used the same system equations as in the original study but changed the semantic context to a cooling system in a nuclear power plant, because we considered this cover story more engaging. Participants took on the role of security managers in charge of reaching and maintaining the optimal temperature in the reactor cooling system of a nuclear power plant to prevent a meltdown. They were told that, whereas they would normally have to supervise the correct functioning of a modern automatic cooling system, their current objective was to get acquainted with the manual handling of a backup cooling system to ensure system stability in case the automatic temperature control failed. In an exploration phase consisting of 20 trials, participants then could enter input values between 0 and 200 to adjust the handwheel, which was the input device of the spare cooling system. The temperature output resulting from such a quantitative control judgment appeared immediately on the screen.

The exploration phase served the sole purpose of familiarizing participants with the experimental task. Therefore, there was no information yet about the aspired level of the outcome variable. Figure 1 shows a stylized version of the input screen used in the nuclear power plant task.

The instruction for the test phases stated that a failure of the automatic cooling system had occurred. This had led to a continuous rise of temperature and a discrepancy between the current coolant temperature and the optimal state of 4°C. Participants were instructed to make quantitative decisions about values of the input device with a possible range from 0 to 200. In each phase, participants worked on the system for a series of 60 trials, with the objective to reach the optimal temperature as soon as possible and then to maintain this optimum. The initial temperature of the coolant was 14°C (i.e., the system was already running hot). In Experiment 1, all participants worked on the control task twice, whereas in Experiment 3 participants faced three test phases. The system was reset before each phase.

At first glance, it might seem easy to control a system with only one input variable and one output variable. However, three features make this task rather difficult. First, the relationship between the system variables is nonlinear, and this nonlinearity is not obvious; it can only be inferred by systematic observation and manipulation of the system over a number of trials. Second, there is a time delay between changes in the input variable and the corresponding reactions of the system. Such a time delay in combination with the initial values of the system parameters causes a sinusoidal oscillation of temperature values. Third, the system has its own dynamics, and without external control judgments (which correspond to a constant setting of the default input value of 100), the system would oscillate and converge toward the suboptimal temperature equilibrium of 11.5°C. The underlying mathematical equations and system parameters are outlined in Appendix A.

We assessed system control performance by calculating the mean absolute deviation (MAD) of control outcomes from the optimal target-value across all trials of a phase save the first. We excluded the first trial when computing the MAD scores because, in this trial, the outcome was not yet influenced by participants' control decisions and, thus, equal for all participants.



The current temperature in the cooling system is **17.25°C.**
The current handwheel value is 80.

Trial 4 out of 60.

New handwheel value:
New temperature:

Figure 1. Stylized input screen for participants working on the nuclear power plant task in Experiments 1 and 3. Participants can enter the desired system input in the white field with the bold borders. As soon as they hit the OK button, the program returns the new system temperature value in the gray box below the input box.

We pretested the power plant control task ($N = 30$ participants) to ensure that it fulfilled an important criterion for the study of group-to-individual transfer and differential weighting, namely relatively stable differences in expertise at controlling the system. Absent any differences in expertise, there is neither learning from the groups' expert nor assigning greater weight to this expert when making a joint control decision. Pretest participants controlled the system individually in two consecutive phases (the system was reset before the second phase). We found a significant positive rank order correlation of participants' control performance in the two phases (Spearman's $\rho = .56$, $p = .001$). In other words, pretest participants who performed better in the first phase also tended to perform better in the second phase, which indicated somewhat stable differences in individual expertise.

Experiment 2. The dynamic control task of Experiment 2 was an inventory management task adapted from Diehl and Sterman (1995). This task is a simplified version of the above-mentioned TAILORSHOP scenario. We chose this task for two reasons. First, using a second task allows us to infer whether potential differences between groups and individuals are task-specific or likely to generalize to other system control tasks. Second, the underlying dynamic system has more variables and is, thus, arguably more difficult than the one used in Experiments 1 and 3, as we will explain after describing the task. Hence, using the second task allows us to test whether groups would benefit from synergy particularly in difficult system control tasks. In the inventory management task, participants took on the role of a production manager in a factory. Here, they had to control the stock inventory by making decisions about the production rate in the face of varying sales. The company sold their products to the public (external sales) but also to their own employees (internal sales). External sales constitute the system's own dynamic. They followed a random walk with a starting value of 400 units in Trial 1 and trial-wise changes drawn from a uniform distribution of integers ranging from −15 to 15 units (this random walk was generated once and was identical for all participants and phases). Internal sales were a positive function of the company's production rate and amounted to 30% of the respective trial's production.

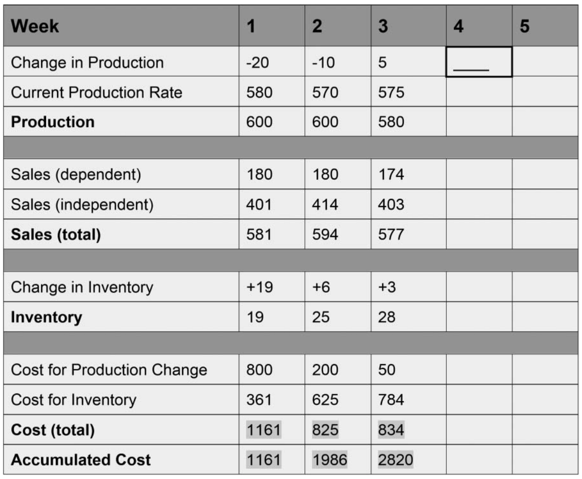
Participants' objective was to minimize the company's total costs. One cost factor, which we labeled inventory costs, was the discrepancy between the inventory at the beginning of a trial and that trial's sales. In case of overproduction (inventory exceeds demand) these costs represent the cost of storage, whereas in the opposite case they represent the costs associated with being unable to deliver products on time. The second cost factor, labeled production costs, comprised the costs of changing the production rates (e.g., hiring or firing labor). As in Experiment 1, participants made one control decision per trial by entering the desired change of the production rate. Production time was two trials resulting in a time-delayed reaction to changes in production rates.

At the end of each trial, participants received information about the following system variables: first, the current production rate as well as the change the participant (or group) had issued in this round; second, the total sales split into external and internal sales; third, the current inventory as well as the change in inventory from the last to the current trial; fourth, the total costs incurred in that trial, split into inventory and production costs; fifth, and finally, the cumulated costs (see Figure 2 for a stylized version of the input screen).

The task setting with a computer interface was comparable to the cooling system scenario with one extension: Participants received a printout of written instructions about the task, explaining the system variables and describing several possible relations between variables. For example, they read that increasing the production rate could either decrease internal demand, leave it unchanged, or increase it, and that it was unclear which of these alternatives was true for the fictional company. Participants could keep the instructions next to their laptops throughout the experiment to allow for a basic understanding of the somewhat complex business context and an awareness of multiple possible alternatives of relations between variables. Again, participants were not informed about specific causal relationships implemented in the system. The underlying mathematical equations of the system are outlined in Appendix B. Participants worked on the system twice, with the system being reset at the beginning of each phase.

The major challenges that the participants faced in the stock inventory task correspond to those of the cooling system scenario. Causal links and strengths of relationships between system variables were not obvious to participants and had to be explored by systematic manipulation and observation of the system. Although system relations are nonlinear and time-delayed in both tasks, we consider the task used in Experiment 2 to be more difficult because it has more system variables (10 vs. 6), the system output is more complex (compare Figures 1 and 2), and participants needed to control two outcomes variables (two cost components) instead of one. Table 1 summarizes the characteristics of the two systems for a direct comparison.

The measure of system control performance was the log2-transformed cumulated costs (i.e., the sum of the two cost components in the system output) at the end of each phase, similar to the original study introducing this task (Diehl & Sterman, 1995). A pretest ($N = 32$), in which participants completed two phases of the stock inventory management task individually (again, the system was reset before the second phase), revealed a significant positive rank order correlation of participants' first and second system control performance (Spearman's $\rho = 0.82$, $p < .001$). Again, this correlation indicates stable differences in expertise, which are a necessary precondition for G-I transfer and effective differential weighting.



Week	**1**	**2**	**3**	**4**	**5**
Change in Production	-20	-10	5		
Current Production Rate	580	570	575		
Production	600	600	580		
Sales (dependent)	180	180	174		
Sales (independent)	401	414	403		
Sales (total)	581	594	577		
Change in Inventory	+19	+6	+3		
Inventory	19	25	28		
Cost for Production Change	800	200	50		
Cost for Inventory	361	625	784		
Cost (total)	1161	825	834		
Accumulated Cost	1161	1986	2820		

Figure 2. Stylized version of the input screen of the inventory management task. Participants can enter the desired change in production rate in the box with the bold borders. As soon as they enter a value, the remaining information for the respective week is displayed.

Procedure

Participants were invited to the laboratory to work on a dynamic system control task on the computer. We ran the experiments using the computer software Assembly Lab for Experiments (ALEX; Schlemmer, 2009). Up to 12 participants were invited at a time and were seated at individual workstations. They were informed about the experiment consisting of a brief exploration phase, an individual test phase (Phase 1), and the assignment to one of the three experimental conditions for the second test phase (Phase 2). In Experiment 3, participants also learned that there was a third test phase, in which all participants would work individually again (Phase 3). We also informed participants about their payment, which consisted of a fixed participation fee and a performance-based bonus. In Experiments 1 and 2, the participation fee was €5, and the bonus payment ranged from €0 to €5 based on Phase 2 performance. In Experiment 3, we increased the fixed fee to €8 because of the longer duration of the experiment. Participants were informed that they could earn a bonus of up to €5 based on their (or their group's) Phase 2 performance, and another €5 based on their Phase 3 performance (for a maximum bonus of €10). In Experiments 1 and 3, we explained to participants that their Phase 2 performance would be evaluated based on the mean absolute deviation from the target value across all 60 trials, while in Experiment 2 they were informed that their bonus was contingent on the cumulative total costs over all 40 trials.[3] The experimenter pointed out that performing well in Phase 1 would allow for better performance in the subsequent phase(s) and, thus, increase the likelihood of receiving a bonus payment. Finally, participants were told that all instructions, further information, and system control devices would be presented on the computer screen.

Participants began by answering demographic questions. Specifically, we asked them to indicate their gender, age, field of study, and which semester they were currently in. In Experiment 1, participants also filled in the German version of the NEO-FFI, a 60-item questionnaire assessing the Big 5 (Borkenau & Ostendorf, 1989). We did not administer the NEO-FFI in Experiments 2 and 3 because exploratory analyses did not show systematic correla-

[3] In Experiments 1 and 3, participants received a bonus of €1 for a Phase 2 MAD score lower or equal to 4.80, €3 if the MAD score was equal or lower to 3.24, and €5 if the MAD was equal or lower than 1.40. These thresholds mark the median, 25th percentile, and best performance among the pre-test participants. In Experiment 2, we defined thresholds at cumulative costs of 100,000, 500,000, and 1,000,000 Euro, respectively. These thresholds were loosely based on the performance of the pre-test participants. They correspond to log2-tranformed cumulative costs of 16.61, 18.93, and 19.93, respectively. An analysis of the amount of bonus payment by decision-maker is available as an online supplement (S1 in the online supplemental materials).

Table 1

Overview of the Two Dynamic System Control Tasks Used in Experiments 1 and 3 (Nuclear Power Plant) and Experiment 2 (Inventory Management)

System characteristics	Nuclear power plant task	Inventory management task
Number of system variables	6	10
Number of input variables	1	1
Number of output variables	1	2
Nonlinear relations	Yes	Yes
Time-delayed relations	Yes	Yes
Number of trials per phase	60	40
Performance criterion	System temperature	Total costs
Performance aim	Stabilize temperature at 4°C	Minimize costs
Performance measure	Mean absolute deviation	$\log_2$-tranformed total costs

tions of any of the five personality dimensions with system control performance.

Prior to the exploration phase, participants were introduced to the task setting without a specific goal. After that, they entered input values and received feedback about the resulting system reactions to get used to the computer interface and to explore the functioning of the system over a period of 20 trials. After the exploration phase, participants answered three questions concerning their motivation to explore the system and were asked to describe their assumptions about the underlying structure of the system.

The instruction for Phase 1 asked participants to control the system according to a specific objective over a series of 60 trials (Experiments 1 and 3) or 40 trials (Experiment 2), respectively. Having fewer trials in Experiment 2 ensured that participants spent about the same time in the lab in Experiments 1 and 2. After each input decision, participants received feedback about the resulting changes in system parameters. The experimenters monitored the participants' Phase 1 performance on a server computer. They then assigned participants to three-person groups so that there was a certain level of performance heterogeneity within the group. A certain amount of within-group heterogeneity is necessary for an informative test of synergy in dynamic system control because, in the absence of notable differences in performance, group members have neither the opportunity to learn from an expert nor to weight that expert's input more strongly when making a joint decision (Baumann & Bonner, 2004). In Experiments 1 and 3, we aimed for each group members' Phase 1 MAD score to differ from the Phase 1 MAD scores of the other two members by at least 0.40. In Experiment 2, group composition was based on differences in Phase 1 cumulated costs. We assigned groups so that pairwise discrepancies amounted to at least 500,000 cost units. After formation, the groups were randomly assigned to either the real group or the nominal group condition.

One third of participants were assigned to the individual condition in Phase 2, which was identical to Phase 1 in all respects. Testing the same number of participants in the individual condition and the (nominal) group condition was technically not necessary. However, when comparing real or nominal groups to individuals, there is almost inevitably larger variance in performance of the latter as compared to the former (cf. Einhorn et al., 1977). Averaging the performance scores of three independent individuals prior to the analyses allowed us to reduce this variance heterogeneity to a certain extent.[4]

In the nominal group condition, participants stayed at their individual workplaces. As in Phase 1, they entered an individual control decision in each trial. Nominal group members learned that they and two other participants would work on the same system. Each person would enter an individual system control decision, and a server computer would then compute the mean of the three decisions and treat it as the system input. The resulting mean was displayed on each nominal group member's computer, but they did not learn about the individual decisions of the other two nominal group members. Nominal group members then received feedback about the state of the system that resulted from their joint control decision. Participants did not know the identity of the other nominal group members, nor did they have the opportunity to interact.

In the real group condition, each group was seated at a table in a separate room. On each trial, all group members first made an individual prediscussion control decision by stating privately the value they would enter on the respective trial. Afterward, the group could discuss these individual suggestions and their underlying control strategies without a time limit in order to reach a consensus decision. Upon entering this consensus decision, the group members received feedback on the system state and proceeded with the next trial in the same fashion. The real group discussions were videotaped in Experiment 3.

Phase 3 of Experiment 3 was similar to Phase 1. All participants worked on the same task individually once more, irrespective of whether they had worked individually, as part of a nominal group, or in a real group in Phase 2. In all three experiments, participants filled out a final questionnaire including questions about their motivation to perform well, about current knowledge concerning the system's structure, and a suspicion check. Afterward, participants were paid, thanked, and debriefed.

[4] Alternatively, we could have compared 35 real or nominal groups with 105 individuals in our analyses. Obviously, the respective means in the individual condition are identical in both cases. Aggregating the data across triplets produces slightly smaller estimates of the error variance than using the individual data at the cost of fewer degrees of freedom. The two analyses produce almost identical results, and none of the conclusions we draw here would change if we analyzed the data in the individual condition on the level of participants rather than aggregating on the level of triplets first.

Results

In the results section, we report the analyses by research question rather than by individual experiment, and we complement the experiment-wise analyses with meta-analyses. In addition to the frequentist statistics we used to analyze our data, we also report Bayes factors obtained via the default settings of the *BayesFactor* package for R. The data and code necessary to reproduce the analyses is available online at https://osf.io/e3uky/ (Schultze & Schulz-Hardt, 2020).

Phase 1 Performance

To create real and nominal groups that were heterogeneous regarding their members' individual capabilities, we assigned participants to the group conditions based on their relative performance rather than using full randomization. This bears the risk of creating baseline differences in initial performance, which would prohibit sensible interpretation of Phase 2 performance. Therefore, we began our analyses by investigating Phase 1 performance in an ANOVA with decision-maker (individual vs. nominal group vs. real group) as between-subjects variable. As we would expect, Phase 1 performance did not differ significantly between conditions in Experiment 1, $F(2, 102) = 1.67$, $p = .194$, $\eta^2 = .03$, $BF = 0.34$. The same was true for Experiments 2 and 3, $F(2, 102) = 0.18$, $p = .837$, $\eta^2 = .004$, $BF = 0.10$, and $F(2, 102) = 1.24$, $p = .294$, $\eta^2 = .02$, $BF = 0.24$, respectively. Thus, the way we assigned participants to experimental conditions in order to create heterogeneous groups did not lead to notable differences in initial performance.

Phase 2 Performance

We first ran a series of ANOVAs with decision-maker (individual vs. nominal group vs. real group) as between-subjects variable to test whether Phase 2 performance differed between experimental conditions. This was the case in all three experiments, $F(2, 102) = 7.76$, $p < .001$, $\eta^2 = .13$, $BF = 42.78$ for Experiment 1, $F(2, 102) = 26.70$ $p < .001$, $\eta^2 = .34$, $BF > 100$ for Experiment 2, and $F(2, 102) = 7.81$, $p < .001$, $\eta^2 = .13$, $BF = 43.10$ for Experiment 3 (see Figure 3). Having found consistently that Phase 2 performance differed between conditions, we proceeded with pairwise comparisons to answer our first two research questions.

Real group versus individual performance. We first tested for the superiority of groups over individuals in dynamic system control by comparing their respective Phase 2 performance. In Experiment 1, this comparison yielded a strong performance advantage of interacting groups, as indicated by substantially lower MAD scores ($M = 2.96$, $SD = 1.67$ vs. $M = 4.52$, $SD = 1.39$), $t(65.88) = -4.25$, $p < .001$, $d = -1.02$, $BF > 100$. Groups also outperformed individuals in the inventory management task of Experiment 2, as reflected in lower $\log_2$-transformed cumulated costs ($M = 20.47$, $SD = 2.53$ vs. $M = 24.82$, $SD = 3.20$), $t(64.54) = -6.33$, $p < .001$, $d = -1.51$, $BF > 100$. In Experiment 3, groups also performed better than individuals, again indicated by lower MAD scores ($M = 2.48$, $SD = 1.75$ vs. $M = 3.66$, $SD = 1.13$), $t(58.23) = -3.36$, $p = .001$, $d = -0.80$, $BF = 24.71$. A random-effects meta-analysis across all three experiments yielded an estimated average effect of $d = -1.10$, 95% CI $[-1.50, -0.69]$, $z = -5.33$, $p < .001$ (see Figure 4). These results show that, in a fair comparison, groups perform better in dynamic system control tasks than individuals.

Nominal group versus individual performance. Having found that interacting groups outperform individuals, we next tested whether at least some of this performance advantage might have resulted from the statistical aggregation effect rather than from real group interaction. To this end, we compared nominal group performance with individual performance in Phase 2. The analyses support the notion of a statistical aggregation effect due to

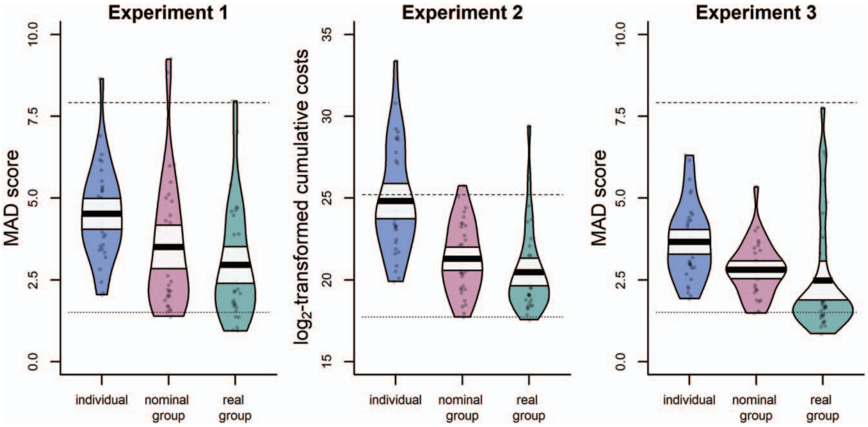


Figure 3. System control performance in Phase 2 by type of decision-maker. Lower numbers indicate better performance. The width of the beans represents the estimated density of the data for a given value of the dependent variable, and they contain the individual data points. The vertical black lines represent the mean, whereas the white bands indicate the respective 95% confidence intervals. The dashed line shows the hypothetical performance that would result from leaving the system unchanged. The dotted line shows the hypothetical performance that would result from entering the input value that eventually stabilizes the system on all trials (Experiments 1 and 3) or always choosing the input value that minimizes costs (Experiment 2). See the online article for the color version of this figure.

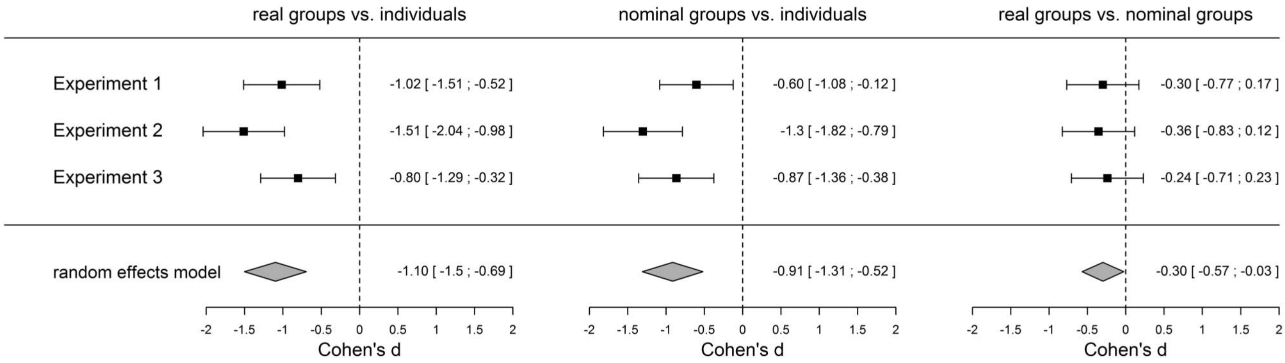


Figure 4. Results of the random-effects meta-analyses of the pairwise comparisons of system control performance between the different types of decision-maker. Lower values indicate better performance. The values to the right of the error bars represent the estimated effect sizes with the respective 95% CIs in brackets.

nominal groups outperforming individuals in all three Experiments. In Experiment 1, MAD scores of nominal groups were significantly lower than those of individuals ($M = 3.51$, $SD = 1.94$ vs. $M = 4.52$, $SD = 1.39$), $t(61.60) = -2.52$, $p = .014$, $d = -0.60$, $BF = 3.50$. Nominal groups also incurred lower cumulated costs in Experiment 2 ($M = 21.29$, $SD = 2.10$ vs. $M = 24.82$, $SD = 3.20$), $t(58.80) = -5.45$, $p < .001$, $d = -1.30$, $BF >$ 100, and lower MAD scores in Experiment 3 ($M = 2.81$, $SD = 0.82$ vs. $M = 3.66$, $SD = 1.13$), $t(61.82) = -3.62$, $p = .001$, $d = -0.87$, $BF = 50.18$. The average effect in a random-effects meta-analysis was large, $d = -0.91$, 95% CI [-1.31, -0.52], $z = -4.53$, $p < .001$ (see Figure 4). These results show that there is a substantial statistical aggregation effect, that is, we can attribute a large part of the superiority of groups over individuals to their superior numbers and statistical aggregation.

Real versus nominal group performance. Having found evidence of both groups' superiority over individuals and a strong statistical aggregation effect, we next address our second research question, namely whether real groups also benefitted from true synergy. To this end, we compared real groups' Phase 2 performance to that of the nominal groups. In Experiment 1, real groups' MAD scores were somewhat lower than those of nominal groups, but this difference was not statistically significant ($M = 2.96$, $SD = 1.67$ vs. $M = 3.51$, $SD = 1.94$)), $t(66.47) = -1.25$, $p = .215$, $d = -0.30$, $BF = 0.48$. The analysis of Experiment 2 produced a similar result. Cumulated costs were lower for real groups, but not significantly so ($M = 20.47$, $SD = 2.53$ vs. $M = 21.29$, $SD = 2.10$), $t(65.86) = -1.49$, $p = .141$, $d = -0.36$, $BF = 0.63$. Finally, we obtained similar results in Experiment 3, with real groups slightly, but not significantly, outperforming nominal groups ($M = 2.48$, $SD = 1.75$ vs. $M = 2.81$, $SD = 0.82$), $t(48.13) = -1.01$, $p = .317$, $d = -0.24$, $BF = 0.38$. Whereas, individually, neither experiment yielded evidence of synergy, the random-effects meta-analysis showed a small and significant synergy effect, $d = -0.30$, 95% CI [-0.57, -0.03], $z = -2.14$, $p =$ .032 (see Figure 4). These results suggest that real groups working on dynamic system control tasks, in fact, benefit from synergy. However, the synergetic effect is small, especially in comparison with the statistical aggregation effect. To quantify the relative size of the statistical aggregation effect, we computed which part of the performance advantage of real groups over individuals we could account for with statistical aggregation. Specifically, we computed the ratio of the performance advantage of nominal groups over individuals to that of real groups. The results suggest that the statistical aggregation effect accounted for 65% (Experiment 1), 81% (Experiment 2), and 72% (Experiment 3) of real groups' superiority over individual decision-makers.

Real group performance versus the average model. Because the meta-analysis comparing real group to nominal group performance yielded evidence of a small synergy effect, we first tested whether this synergy was attributable to groups weighting their more accurate members more when deciding on a joint system input. To this end, we first computed the average of real group members' individual prediscussion decisions. We then computed the performance this averaged value would have yielded given the current state of the system. In other words, this average model yielded the performance that the groups would have achieved if they had simply averaged their members' prediscussion decisions (i.e., equal weighting). Computing this average model as a performance baseline is conceptually similar to computing the group potential as the average input of three individuals, but it already accounts for the fact that real group members may have benefitted from G-I transfer. Finally, we compared real groups' actual system control performance with the average model. In Experiment 1, real groups' MAD scores were slightly lower than those that would have resulted if groups had merely averaged their members' prediscussion decisions ($M = 2.96$, $SD = 1.67$ vs. $M = 3.04$, $SD = 1.60$), but this difference was not significant, $t(34) = -1.06$, $p = .297$, $d = -0.18$, $BF = 0.30$. The corresponding analysis of Experiment 2 produced a similar result, with real groups performing slightly but not significantly better than the average model ($M = 20.47$, $SD = 2.53$ vs. 20.48, $SD = 2.52$), $t(34) = -1.13$, $p = .268$, $d = -0.19$ $BF = 0.32$. The same was true for Experiment 3. Real groups' MAD scores were slightly lower than those that would have results from averaging, but the difference was not significant ($M = 2.48$, $SD = 1.75$ vs. $M = 2.51$, $SD = 1.71$), $t(34) = -0.47$, $p = .643$, $d = -0.08$, $BF = 0.20$. Aggregating the data in a random-effects meta-analysis produced small but insignificant performance differences favoring the real groups, $d = -0.15$, 95% CI [-0.34, 0.05], $z = -1.49$, $p = .137$. In sum, we did not find support for the idea that synergy in dynamic system control stems from differential weighting of group member inputs because groups' performance did not significantly exceed the performance that would have resulted from weighting all members equally.

Phase 3 Performance

We conclude our analysis of system control performance by testing for G-I transfer as a source of synergy. To test for G-I transfer, Experiment 3 included another individual test phase after the group phase. Note that such a postgroup phase is necessary for an unequivocal test of G-I transfer, because effects of statistical aggregation contaminate the individual Phase 2 performance in real and nominal groups. That is, real and nominal group members may perform substantially better individually in Phase 2 not because their understanding of the task improved but because statistical aggregation within the group forces them into more desirable system states. It is also conceivable that real group members benefit from G-I transfer but that they cannot contribute their newly gained system knowledge fully in the subsequent group discussions (because we have no means to test whether this was the case, we do not consider this possibility further). In any case, we can conclude unequivocally that G-I transfer occurred if, in the postgroup phase, participants in the real group condition show better individual control performance than participants in the other two conditions.

As mentioned above, it is theoretically possible that nominal group members could also benefit from G-I transfer because they learned about the mean of their own and the other two nominal group members' input decision. Therefore, we cannot rule out that former nominal group members might perform somewhat better individually than participants who worked as individuals in Phase 2—in case this happens, this would make our test for G-I transfer even more conservative, because it would then be harder for real group members to outperform nominal group members in Phase 3. Importantly, there are no beneficial effects of statistical aggregation in either group condition, because all participants worked individually in Phase 3. Accordingly, we might also expect the performance differences between types of decision-makers to be somewhat stronger in Phase 3 than in Phase 2. Because the data of participants who were members of real or nominal groups in Phase 2 were not independent, we, again, analyzed performance on the group level.

An ANOVA on Phase 3 performance with decision-makers as the independent variable showed significant differences between the experimental conditions, $F(2, 102) = 8.12$, $p < .001$, $\eta^2 = .14$, $BF = 54.54$. We followed this analysis with pairwise comparisons. Participants who were part of a real group in Phase 2 outperformed those who worked as individuals previously, as indicated by significantly lower MAD scores ($M = 2.43$, $SD = 1.34$ vs. $M = 3.35$, $SD = 1.27$), $t(67.81) = -2.96$, $p = .004$, $d = -0.71$, $BF = 9.12$. We did not observe similar performance advantages among former nominal group members. Descriptively, their Phase 3 performance was even slightly worse than that of individuals, but this difference was not statistically significant ($M = 3.51$, $SD = 1.01$ vs. $M = 3.35$, $SD = 1.27$), $t(64.86) = 0.60$, $p = .550$, $d = 0.14$, $BF = 0.28$. Most importantly, former real group members performed better than former nominal group members in Phase 3 ($M = 2.43$, $SD = 1.34$ vs. $M = 3.51$, $SD = 1.01$), $t(63.38) = -3.83$, $p < .001$, $d = -0.91$, $BF = 89.57$. This finding suggests that participants who worked in real groups in Phase 2 benefitted from G-I transfer (see Figure 5).

Exploratory Analyses

Comparing real group performance to various models. We begin our exploratory analyses with a number of model comparisons to gain a better insight into how well groups performed when combining their members' individual prediscussion control decisions into a joint group decision. We already know from our confirmatory analyses that group performance was slightly but not significantly better than what would have resulted from simply averaging the three group members' individual decisions. Here, we compare real group performance with four additional models. The first is the median model, which assumes that groups always take the median of their members' individual decisions, thus discounting outliers. The second model assumes that groups always take the individual decision of the group member who had performed best in Phase 1 (initial best member model), whereas the third assumes they always take the individual decision of the group member who performed best individually in Phase 2 (current best member model). The two best member models differ for two reasons: first, because G-I transfer allows weaker members to improve in individual accuracy with a certain chance of becoming the group's expert and, second, because—even in the absence of G-I transfer—the relative performance within groups may change due to regression to the mean, or due to other factors. As such, the current best member model is a more challenging performance benchmark than the initial best member model. Finally, we compared group performance with the accuracy model, which always takes the individual decision that would have led to the best outcome irrespective of which member made it. For all model comparisons, we contrasted real groups' MAD scores or log2-transformed cumulative costs, respectively, with those obtained via the different models, such that negative values indicate an

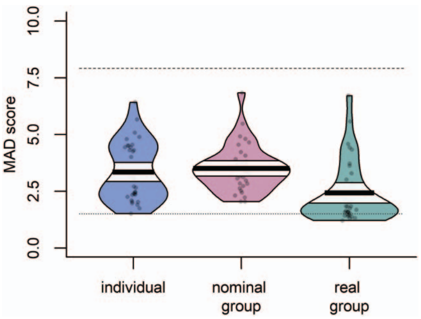


Figure 5. Individual system control performance in Phase 3 of Experiment 3 by type of decision-maker in Phase 2. Lower scores indicate better performance. The width of the beans represents the estimated density of the data for a given value of the dependent variable, and they contain the individual data points. The vertical black lines represent the mean, whereas the white bands indicate the respective 95% confidence intervals. The dashed line shows the hypothetical performance that would result from leaving the system unchanged. The dotted line shows the hypothetical performance that would result from entering the input value that eventually stabilizes the system on all trials. See the online article for the color version of this figure.

advantage of the respective model whereas positive values indicate that the groups performed better.

As shown in Figure 6, real groups' Phase 2 performance did not differ significantly from the median model in either experiment, all $|t(34)| < 1.26$, all $p > .219$, all $|d| < 0.22$, all $BF < 0.38$, meta-analytic effect size estimate $d = 0.13$, 95% CI $[-0.06, 0.33]$, $z = 1.37$, $p = .171$. The same was true for the initial best member model, all $|t(34)| < 1.81$, all $p > .079$, all $|d| < 0.31$, all $BF < 0.78$, meta-analytic effect size estimate $d = 0.07$, 95% CI $[-0.29, 0.43]$, $z = 0.37$, $p = .711$. In contrast, real group performance fell short of the current best member model in Experiments 1 and 3, both $t(34) < -3.61$, both $p < .001$, both $d < -0.61$, both $BF > 32$. In Experiment 2, the current best member model also outperformed real groups, but not significantly so, $t(34) = -1.45$, $p = .156$, $d = -0.25$, $BF = 0.47$. The meta-analysis yielded a clear performance advantage of the current best member model, $d = -0.56$, 95% CI $[-0.93, -0.20]$, $z = -3.02$, $p = .003$. Finally, the groups performed substantially worse than the accuracy model in all three experiments, all $t(34) < -4.69$, all $p < .001$, all $d < -0.79$, all $BF > 100$, meta-analytic effect size estimate $d = -1.08$, 95% CI $[-1.41, -0.75]$, $z = -6.36$, $p < .001$.

Comparison of real and nominal group performance over of time. In the next set of exploratory analyses, we investigated how the Phase 2 performance of real as compared to nominal groups developed over time (Figure 7 shows Phase 2 performance by trial for real and nominal groups and, for the sake of completeness, of individuals). The reason for this analysis was that the performance advantage of real groups was quite small and only detectable in the meta-analysis. Since the results of Experiment 3 suggest G-I transfer to produce this performance advantage, and because G-I transfer might take some time to manifest, we tested the possibility that real groups might outperform nominal groups over time with more pronounced performance differences at later stages of Phase 2. We z-standardized real and nominal group performance in Phase 2, formed 10 trial blocks in each experiment, and aggregated the performance scores in each trial block. In Experiments 1 and 3, each trial block contained six trials with the exception of Block 1, which contained only five trials because performance in the cooling system scenario used in these two experiments could not yet differ on the first trial. In Experiment 2, each trial block consisted of four trials. We ran mixed ANOVAs with decision-maker (real group vs. nominal group) as a between-subjects variable, trial block (1 to 10) as a within-subjects variable, and group performance as the dependent variable.[5] The results were consistent across experiments. First, there were main effects of trial block in all three experiments, all $F(9, 612) > 26.68$, all $p < .001$, all $\eta^2 > .16$, all $BF > 100$. Second, the effect of decision-maker was not significant in either Experiment, all $F(1, 68) < 2.63$, all $p > .108$, all $\eta^2 < .02$, all $BF < 19.80$.[6] Finally, there was no evidence of an interaction effect, all $F(9, 612) < 1.81$, all $p > .064$, all $\eta^2 < .02$, all $BF < 0.02$. As a meta-analysis, we ran a linear mixed model with decision-maker, trial block, and their interaction as fixed effects and random intercepts for groups nested within experiments. Note that this amounts to a fixed effects meta-analysis of the main and interaction effects (models containing random slopes of these effects did not converge). Consistent with the confirmatory meta-analysis of the aggregate Phase 2 performance, we found an effect of decision-maker due to better performance in the real groups as compared to the nominal groups, $F(1, 206) = 5.33$, $p = .022$. There was also an effect of trial block, $F(9, 1872) = 31.49$, $p < .001$, but this effect is difficult to interpret because the direction of the time effect differed between Experiment 1 and 3, on the one hand, and Experiment 2, on the other (see Figure 7). Whereas deviations from the optimal temperature decreased over time in Experiment 1 and 3, cumulated costs necessarily increased over time in Experiment 2, and these trajectories are also reflected in the standardized performance scores. The meta-analytic interaction effect was not significant $F(9, 1872) = 0.54$, $p = .846$. In sum, these analyses confirm a small performance advantage of real groups, but they also suggest that this small advantage manifests quite early and does not change substantially over the course of the experiments.

Analysis of individual system control decisions in Phase 2. We can test the idea that the performance advantage of real groups over nominal groups is due to G-I transfer further by investigating the individual control decisions in Phase 2. If real groups benefitted from G-I transfer, we should expect that (a) real group members' individual decisions in Phase 2 are more similar to each other (i.e., less heterogeneous) than those of both nominal group members or triplets of individuals and that (b) real group members' individual Phase 2 performance exceeds that of nominal group members and individuals. We ran two exploratory analyses to test whether this was the case.

Concerning heterogeneity within groups (or within triplets of individuals), we first built a heterogeneity index by computing the standard deviation of the three individual system control decisions for each trial and then averaging these standard deviations across all trials. We then analyzed the heterogeneity of the individual decisions in ANOVAs with decision-maker (individual vs. nominal group vs. interacting group) as between-subjects variable. The effect of decision-maker was significant in all three experiments, all $F(2, 102) > 11.31$, all $p < .001$, all $\eta^2 > 0.18$, all $BF > 100$. As shown in Figure 8, the differences in heterogeneity were largely driven by lower heterogeneity in real groups compared to the other two conditions. A random-effects meta-analysis of the pairwise comparisons confirmed this impression. Real group members' individual decisions were less heterogeneous than those of both individuals, $d = -1.96$, 95% CI $[-2.92, -1.00]$, $z = -3.99$, $p < .001$, and nominal group members, $d = -1.69$, 95% CI $[-2.33, -1.05]$, $z = -5.14$, $p < .001$. Nominal group members' individual decisions were also somewhat less heterogeneous than those of the individuals, $d = -0.41$, 95% CI $[-0.71, -0.11]$, $z = -2.71$, $p = .007$. However, given that there was no evidence of G-I transfer in the nominal groups in Experiment 3, it seems plausible that the lower heterogeneity in the nominal groups resulted from nominal group members being forced into more desirable system states due to statistical aggregation.

[5] We computed the Bayes factors for the three effects as follows. For the main effects, we compared a model containing only the respective main effect to the null model. For the interaction effect, we compared the full model to a model containing only the two main effects.

[6] The BF for the main effect of condition was 19.79 in Experiment 2 despite the nonsignificant results in the frequentist analysis. More in line with the frequentist analysis, it was 1.56 in Experiment 1 and 0.35 in Experiment 3.

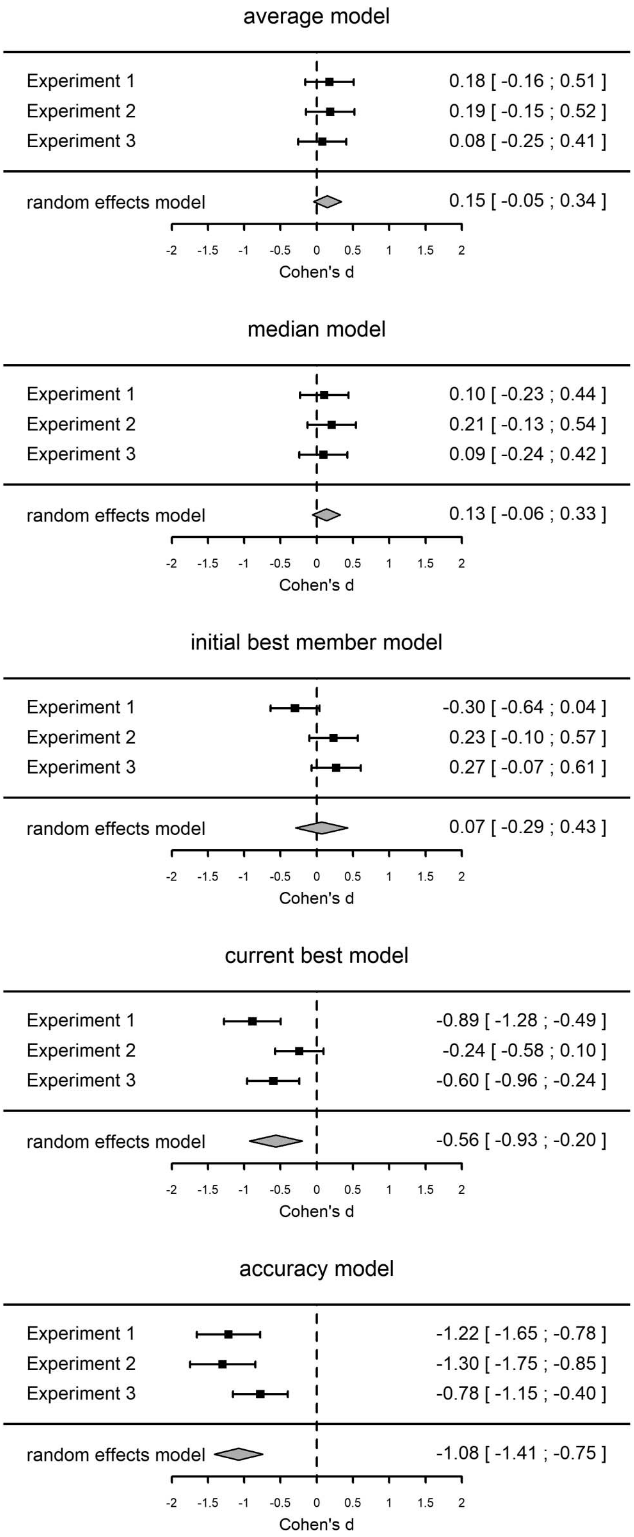


Figure 6. Comparison of real groups' Phase 2 performance to different models combining the group members' individual prediscussion decisions into a hypothetical group decision. Negative values indicate that the respective model performed better than the real groups (i.e., the model produced lower deviations from the target value or lower cumulated costs). The values to the right represent the estimated effect sizes with the respective 95% confidence intervals in brackets.

With regard to individual Phase 2 performance, we first computed the hypothetical individual Phase 2 performance of real and nominal group members because their individual prediscussion decisions were not actually implemented unless they were identical to the group decisions (real groups) or to the mean of the three individual decisions (nominal groups). To this end, we computed the system state that would have resulted from entering the respective group members' individual control decision and used this hypothetical system state to compute the respective performance measures. We averaged the *z*-standardized performance scores of each participant across all trials and, then, aggregated these mean performance scores within each real or nominal group and within each triplet of individuals to account for dependencies in the real and nominal group data (see Figure 9). We analyzed individual Phase 2 performance in the same fashion as the within-group heterogeneity. The respective ANOVAs yielded significant differences in individual performance between conditions, all $F(2, 102) > 7.30$, all $p < .002$, all $\eta^2 > 0.12$, all $BF > 29$. The meta-analytic pairwise comparisons showed that real group members' individual Phase 2 performance was superior to that of both individuals, $d = -1.13$, 95% CI $[-1.61, -0.65]$, $z = -4.58$, $p < .001$, and nominal group members, $d = -0.40$, 95% CI $[-0.67, -0.13]$, $z = -2.88$, $p = .004$. Nominal group members' individual Phase 2 performance also exceeded that of the individuals, $d = -0.82$, 95% CI $[-1.24, -0.40]$, $z = -3.83$, $p < .001$, but, as with the lower heterogeneity, this effect most likely reflects the benefits of statistical aggregation rather than G-I transfer.

Improvement in individual performance by relative expertise. In this exploratory analysis, we investigated how individual performance in Experiment 3 increased between Phases 1 and 3 dependent on the relative performance of a participant within their real or nominal group or triplet of individuals. This analysis can provide further evidence of G-I transfer. Specifically, if there was G-I transfer, we would not only expect that individual performance increases more in the real groups, but that these increases are stronger for the real groups' weaker members, because they are the ones who can learn from the best members. We tested this idea in a mixed ANOVA with decision-maker in Phase 2 (real group vs. nominal group vs. individual) as a between-subjects factor and group member (best vs. medium vs. worst) as a within-subjects factor. As the dependent variable, we computed performance increases as the difference between participants' Phase 1 MAD scores and Phase 3 MAD scores such that positive values indicate performance increases. The ANOVA revealed significant effects of decision-maker, $F(2, 102) = 7.93$, $p < .001$, $\eta^2 = .06$, $BF = 19.46$, group member, $F(2, 204) = 94.58$, $p < .001$, $\eta^2 = .27$, $BF > 100$, and their interaction, $F(2, 204) = 7.96$, $p < .001$, $\eta^2 = .08$, $BF > 100$ (see Figure 10).

To disentangle the interaction, we analyzed performance improvements separately for each type of group member. On average, the best members did not improve significantly between Phases 1 and 3, $t(104) = 1.27$, $p = .205$, $d = 0.12$, $BF = 0.24$, and there were also no differences in best members' performance increases between conditions, $F(2, 102) = 1.28$, $p = .284$, $\eta^2 = .02$, $BF = 0.25$. The performance of the respective medium members did, on average, increase signifi-

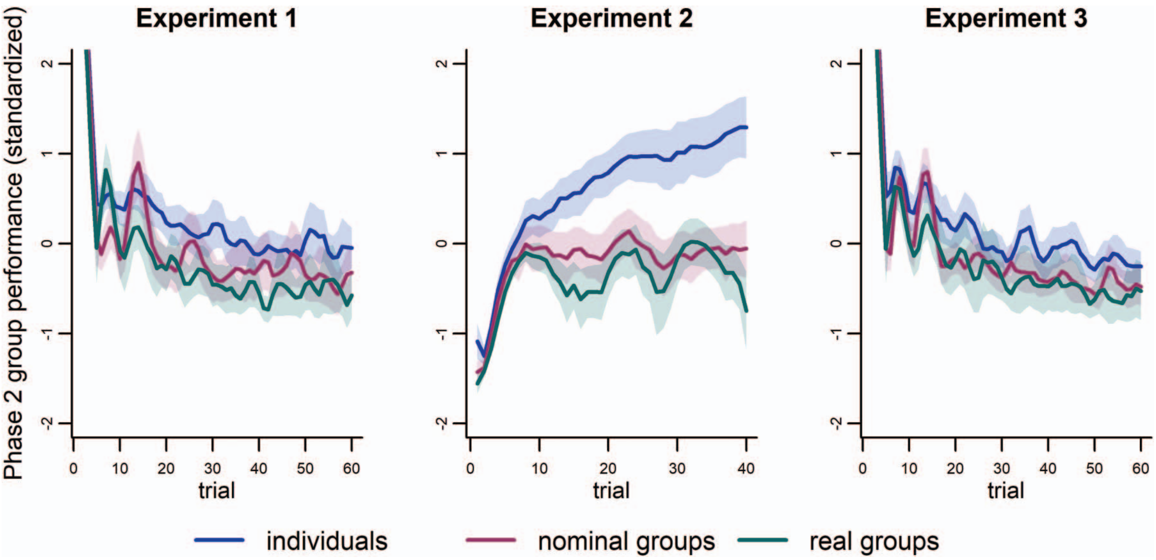


Figure 7. Phase 2 performance (*z*-standardized) across time by type of decision-maker. Lower scores indicate better performance. The bold lines represent the means, whereas the semitransparent bands indicate the corresponding 95% confidence intervals. Performance in the individual condition is computed as the average performance of triplets of individuals. Note that the statistical comparison of real and nominal group performance is based on 10 trial blocks instead of all individual trials. See the online article for the color version of this figure.

cantly, $t(104) = 5.25$, $p < .001$, $d = 0.51$, $BF > 100$, but again, there were no significant differences in these performance increases between conditions, $F(2, 102) = 2.54$, $p = .084$, $\eta^2 = .05$, $BF = 0.69$. Worst members' performance also increased between Phases 1 and 3, on average, $t(104) = 10.94$, $p < .001$, $d = 1.07$, $BF > 100$. However, performance increases of the respective worst members differed strongly between conditions, $F(2, 102) = 12.87$, $p < .001$, $\eta^2 = .20$, $BF > 100$. Pairwise comparisons showed that the individual performance of former real group members improved more than that of both former nominal group members and individuals, $t(54.78) = 3.10$, $p = .003$, $d = 0.74$, $BF = 12.93$, and $t(64.00) = 5.93$, $p < .001$, $d = 1.42$, $BF > 100$, respectively. In contrast, performance increases did not differ significantly between former nominal group members and individuals, $t(104) = -1.65$, $p = .104$, $d = -0.39$, $BF = 0.78$. In sum, these results support the notion that real group members, but not nominal group members, benefitted from G-I transfer.

Analysis of group discussions. Because we videotaped the group discussions in Experiment 3, we ran number of exploratory analyses to gain some insights into the processes underlying G-I transfer. The general approach we used here was to predict the magnitude of G-I transfer from certain aspects of the group discussion.[7] We first explored whether group discussion time as an indicator of discussion intensity was correlated to the magnitude of the G-I transfer. However, this was not the case, $r = -.13$, $p = .460$, $BF = 0.47$.

We next tested the relation of the depth of information elaboration and G-I transfer. To this end, we had two independent and hypothesis-blind research assistants code the group discussions for the first 20 trials of Phase 2 (we restricted the video analysis to the first 20 trials because these trials accounted for more than half of the total discussion time). As a measure of information elaboration, we used the same approach as Keck and Tang (2019), who adopted a scale to measure the depth of information elaboration in group discussions from van Ginkel and van Knippenberg (2008). Specifically, the research assistants rated the depth of information elaboration in each of the 20 discussions for each of the 35 groups on a Likert scale ranging from 1 (*no elaboration*) to 7 (*very deep elaboration*). A rating of 1 was used when group members did not mention any information about the system and, instead, only exchanged their prediscussion decisions and/or discussed their joint decision. A score of 7, in contrast, was assigned when groups shared a large amount of information and hypotheses about the system, elaborated on the shared information, and discussed the relations of the system's variables. Agreement between the two raters was high, $r = .84$, $p < .001$, $BF > 100$, so we averaged their ratings. We then computed the average depth of information elaboration per group and correlated it with the magnitude of G-I transfer. While this correlation was in the expended direction, it was small and not statistically significant, $r = .20$, $p = .209$, $BF = 0.74$.

Finally, we turned to the discussion content. Specifically, there were five critical pieces of system knowledge (see Table 2), and

[7] At first, it may seem intuitive to use Phase 2 group performance as the criterion of interest in these analyses. However, using G-I transfer as the criterion is preferable because it follows a causal logic. Group discussions early in Phase 2 can plausibly cause improvements in individual performance between Phases 2 and 3, whereas the opposite is not possible. In contrast, using Phase 2 group performance as the criterion bears the risk of reverse causality. Specifically, groups that performed relatively better in Phase 1 are very likely to perform relatively better in Phase 2 as well, and these well-performing group may engage in more elaborate discussions.

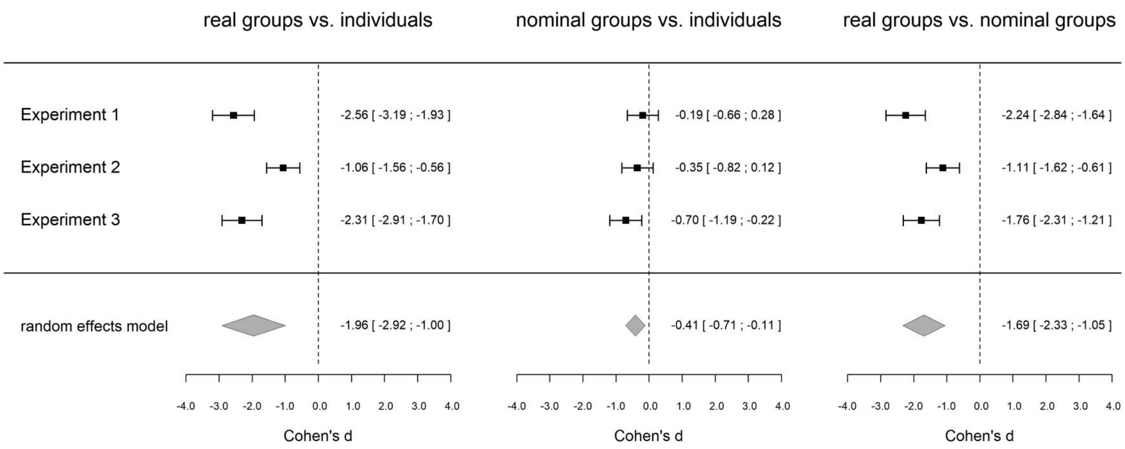


Figure 8. Heterogeneity of individual system control decisions in Phase 2 by type of decision-maker. The bold lines represent the means, whereas the semitransparent bands indicate the corresponding 95% confidence intervals. Note that the statistical analyses are based on the mean across all trials for the sake of simplicity. The extreme spike in heterogeneity within triplets of individuals in Experiment 2 is attributable to one participant increasing the production rate by 15,000 units in Trial 31 and immediately decreasing it by 15,000 units in Trial 32.

we looked at whether groups discussed them during the first 20 trials. To this end, the two research assistants coded during each group discussion whether the group had discussed a piece of system knowledge in the respective trial. Agreement between raters was generally high, but varied somewhat between the pieces of system knowledge, ranging between 81% and 100% of the cases. Therefore, we only coded a piece of system information as discussed by a group if both raters unanimously coded it as discussed in at least one trial. As can be seen in Table 2, none of the groups discussed the nuisance effect of the outside temperature, which is why we do not consider it further in the analysis. We used discussion of the remaining four pieces of system knowledge to predict the magnitude of G-I transfer in a linear regression. As predictors, we included four dummy variables, each coding whether a group had discussed the respective piece of system knowledge.[8] Discussion of the time delay and discussion of the direction of the input-output relationship were significant predictors of the magnitude of G-I transfer, $\beta = .39$, $p = .016$, $BF = 4.31$, and $\beta = .34$, $p = .031$, $BF = 2.57$, respectively. In contrast, neither discussion of the systems' own dynamic nor discussion of the optimal input levels were significantly related to G-I transfer, $\beta = -.08$, $p = .513$, BF = 0.38, and $\beta = .21$, $p = .117$, $BF = 0.96$, respectively. In total, the regression model accounted for 54% of the variance in G-I transfer, suggesting that discussion of system knowledge is critical for G-I transfer in dynamic system control tasks.

Discussion

Successful control of complex dynamic systems like cooling systems in nuclear power plants or managing production and inventory in companies is highly desirable to maintain public safety and prosperity. Previous research has shown that individuals often fare poorly when controlling dynamic systems of a certain complexity. One pressing, but still unanswered, question is whether employing groups can lead to superior system control performance. The present study, which included three experiments and used two different scenarios (a cooling system in a nuclear power plant and a stock inventory management task in a business setting), aimed to provide an unequivocal answer to this question. We further tested whether groups working on dynamic system control tasks benefit from true synergy beyond the statistical aggregation effect (i.e., the beneficial effects of statistical aggregation of group members' individual inputs). Finally, our objective was to identify the mechanisms underlying synergy in dynamic system control. To these ends, we introduced a novel methodological approach comparing the performance of interacting groups not only with individuals but also with noninteracting nominal groups working on a joint system. In our experiments, nominal groups benefitted from the statistical aggregation effect, but since their members could not interact, they could not reap the potentially synergetic effects of within-group interaction.

The results of the three experiments provide the first unequivocal evidence that groups outperform individuals when controlling dynamic systems. The lion's share of this performance advantage stems from the statistical aggregation of independent opinions, as indicated by the superiority of nominal groups over individuals. However, the meta-analytic comparison of real and nominal group performance also showed that within-group interaction led to a small (but significant) increase in performance; that is, real groups benefitted from true synergy. This finding is particularly noteworthy given that previous research on small group performance often failed to find benefits of social interaction (see Kerr & Tindale, 2004), which led Armstrong (2006) to question whether there was ever a benefit to having decision-making groups interact. As such, our findings add to a growing literature suggesting that Armstrong's

[8] We computed a Bayes factor for each of the four predictors by comparing the full model containing all four predictors against a model lacking the predictor of interest. Thus, the Bayes factor indicates that predictor's unique contribution.

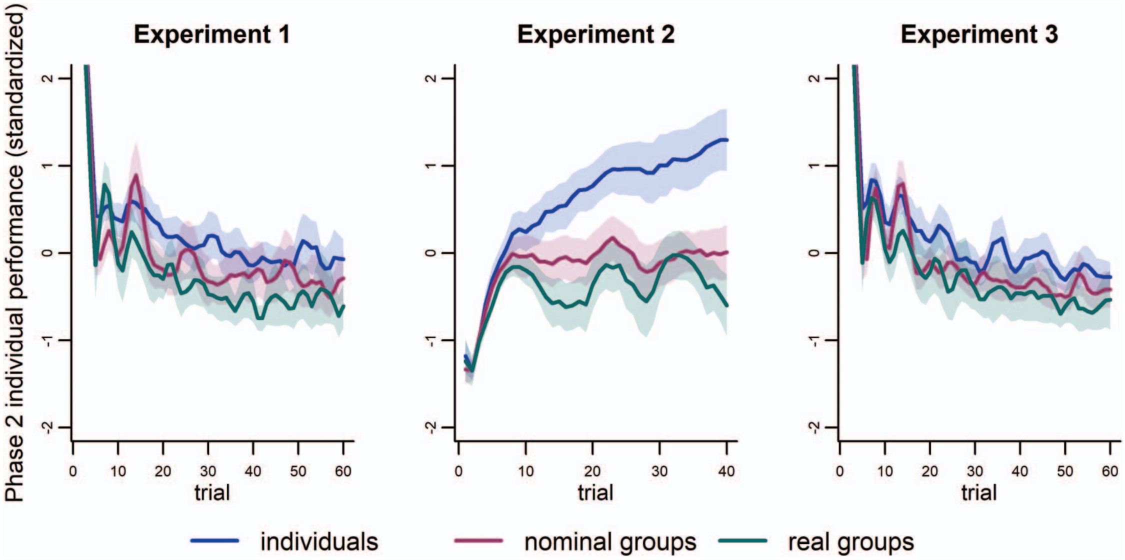


Figure 9. Individual performance (*z*-standardized) in Phase 2 by type of decision-maker. The bold lines represent the means, whereas the semitransparent bands indicate the corresponding 95% confidence intervals. Note that the statistical analyses are based on the mean across all trials for the sake of simplicity. See the online article for the color version of this figure.

call to abolish group discussion as a decision-making tool might be premature. We contribute to this literature by identifying another important context in which groups benefit from interaction. Whereas previous research has shown synergy in a number of static problems such as quantitative judgment (Minson et al., 2018; Schultze et al., 2012) or simple problem solving (Laughlin et al., 2003; Laughlin et al., 2006), our study shows that similar effects also occur in dynamic problems.

The Sources of Synergy in Dynamic System Control Tasks

We tested two possible mechanisms underlying the observed synergy: differential weighting of group members' individual inputs and G-I transfer. We found no evidence of differential weighting in any of the three experiments. On average, the performance of real groups did not differ substantially from a simple average of the suggestions their members had made individually on each trial prior to discussion. However, there was clear evidence of G-I transfer. In Experiment 3, we added an individual test phase after the group phase (or nominal group or individual phase, respectively). Here, participants who had previously worked in real groups showed better individual performance than both participants who had worked in nominal groups and participants who had worked as individuals before. In contrast, individual performance of former nominal group members and individuals did not differ significantly. In other

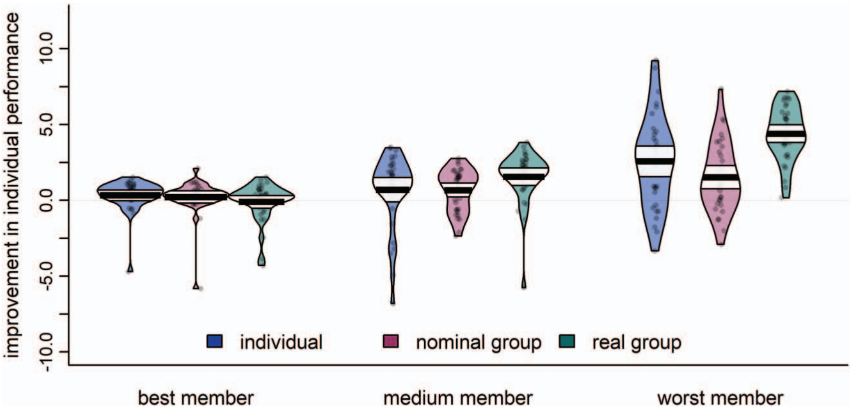


Figure 10. Improvement in individual system control performance between Phases 1 and 3 by type of type of decision-maker in Phase 2 and participants' relative expertise within their group or triplet of individuals. The width of the beans represents the estimated density of the data for a given value of the dependent variable, and they contain the individual data points. The vertical black lines represent the mean, whereas the white bands indicate the respective 95% confidence intervals. See the online article for the color version of this figure.

Table 2
Discussion of System Knowledge in Real Groups

Piece of system knowledge	Discussed within the first 20 trials		Agreement between raters
	Yes	No	
Time delay of input–output relation	27	8	91%
Oscillation (own dynamic)	1	34	100%
Direction of input–output relation	32	3	81%
Optimal input levels (±10 points)	13	22	97%
Nuisance effect of outside temperature	0	35	100%

Note. Discussions of time delay and oscillation were coded as dichotomous variables (discussed vs. not discussed). The remaining three pieces of system knowledge were coded using trichotomous variables (discussed correctly vs. discussed incorrectly vs. not discussed). Here, we only consider system knowledge as being discussed if it was discussed correctly at least once during the first 20 trials of Phase 2.

words, group interaction lead to better performance beyond simple statistical aggregation, because group members learned to control the system more effectively during the group discussions. Our exploratory analyses further showed that, in line with the idea of G-I transfer, individual Phase 2 decisions were more similar to each other within real groups than within nominal groups or triplets of individuals, and they were more accurate, too. In addition, performance increases attributable to social interaction predominantly affected the weaker group members. Finally, increases in real group members' individual performance were greater when groups discussed critical system knowledge such as time-delays or the direction of the input-output relations. Thus, we cannot only state that real groups benefitted from G-I transfer but also pinpoint the improvements in real group members' ability to perform the task to the socially transmitted acquisition of causal system knowledge.

Our findings complement earlier studies highlighting the relevance of G-I transfer for the emergence of synergy (Schultze et al., 2012; Stern et al., 2017). Although these studies used simple quantitative judgment tasks instead of dynamic systems, we note some similarities in the results. In quantitative judgment tasks, a single group interaction is often sufficient to induce the complete G-I transfer; that is, further group discussions often do not lead to further improvements in individual performance (Stern et al., 2017). As our exploratory analyses of individual Phase 2 performance suggest, G-I transfer manifests quickly in dynamic system control tasks, too, taking at best a few group interactions. Consider, however, that discussion times were not limited, and that groups spent considerably longer discussing the task in earlier trials. As such, G-I transfer manifesting early during the group phase or—in case of previous studies—after a single interaction might be the result of group members taking their time to discuss the problem and exchange task-relevant information when they meet for the first time. Another similarity to previous research on G-I transfer in simple quantitative judgment tasks is that G-I transfer only benefitted the groups' weaker members whereas the performance of the strongest members did not improve (Schultze et al., 2012). This observation fits the idea that the groups' experts share their task-relevant knowledge with the other group members who, in turn, become better at performing the task.

Despite these similarities, we argue that the content of learning must necessarily differ. In the context of simple quantitative estimation tasks, G-I transfer mainly consists of group members correcting their erroneous metric knowledge, that is, they learn sensible scales for their estimates during discussion (Stern et al., 2017). For example, participants who estimate the length of rivers may learn that the longest river on Earth is close to 7,000 km in length and, after learning this fact, drastically lower their subsequent estimates. Accordingly, learning a reasonable frame of reference in a discussion is sufficient to increase performance in quantitative judgment tasks (Bonner et al., 2007). In our study, in contrast, scaling effects cannot account for G-I transfer. We observed G-I transfer only when participants had previously worked in interacting groups, but not when they had been part of a nominal group. This is important, because nominal group members were aware of the reference values of averaged individual system input values during the second test phase. In addition, we found that G-I transfer was stronger when groups discussed the nature of the system's variable relations, in particular the direction of the input-output relation and the fact that this relation was time-delayed. Therefore, the data of Experiment 3 suggest that real group members obtained greater causal knowledge of the system, which, in turn, allowed them to perform better individually. This observation is in line with theorizing on dynamic system control tasks, which highlights the necessity to acquire knowledge about the causal structure of a system in order to manage it well (Osman, 2010). It also fits nicely with recent research on synergy in multicue judgment tasks showing that group discussions lead synergy via G-I transfer of knowledge about cue-target relations (Lippold, Schulz-Hardt, & Schultze, in press).

In contrast to G-I transfer, we did not find evidence of synergy resulting from differential weighting of group members' prediscussion decisions. Instead, group performance did not differ significantly from a simple average of these prediscussion decisions. These findings are also in line with previous research testing for effects of differential weighting in quantitative group judgments after controlling for G-I transfer (Schultze et al., 2012; Stern et al., 2017), where differential weighting was only found under very limited conditions.

It is worth mentioning here, though, that finding evidence for differential weighting is not easy for two reasons. First, as stated initially, for groups to benefit from differential weighting, their members must differ in expertise. Although we created groups such that there were notable differences in expertise between group members, G-I transfer tends to reduce these differences, because weaker group members learn from their more knowledgeable counterparts whereas stronger members' performance remains stable. This is exactly what we found in our exploratory analysis of the performance increases between Phases 1 and 3 in Experiment 3. Thus, G-I transfer might have been sufficient to reduce differences in expertise between group members to a level that made benefitting from differential weighting difficult. Second, the two models that outperformed real groups, namely the current best member model and the accuracy model, both capitalize on chance to some extent. For example, a few lucky guesses may allow a less knowledgeable group member to outperform the group member with the most accurate system knowledge in Phase 2 simply by chance, requiring the group to rely more on the lucky member than on the group's expert. The accuracy model amplifies this problem, because it requires groups to recognize with certainty which individual decision is the most accurate even if the member who suggested it admits to purely guessing (Gigone & Hastie, 1997).

Although we did not find any evidence of differential weighting in our experiments, it is conceivable that differential weighting may play a more relevant role in dynamic systems with a greater number of input variables and/or more complex variable relationships. Especially if system knowledge is hard to verbalize and difficult to explain to other group members, individual system knowledge might not easily (or only partially) transfer to other group members. Thus, there might be a smaller effect of G-I transfer in systems that are more complex. If so, we would expect the heterogeneity of group members' individual expertise to persist longer, thus providing groups with the opportunity to benefit from assigning greater weight to their experts.

The Costs and Benefits of Group Work in Dynamic System Control

We have shown that communication and interaction in an interacting group context lead to superior performance when controlling dynamic systems. However, the performance advantage of interacting real groups over the noninteracting nominal groups is relatively small with an average effect size of about $d = 0.30$, and it comes at a cost: Groups required considerably more time to come up with a consensus decision than both individuals and nominal groups (see Supplement S2 in the online supplemental materials, which contains an analysis of discussion times). In practice, bringing people together and having them discuss and coordinate decisions may also imply additional monetary costs. Therefore, practitioners should carefully consider the benefits and costs of collaboration when deciding whether they want to authorize groups or individuals with the control of a dynamic system.

If performance is critical and there are no time constraints, interacting groups should prove especially useful because of their superior performance over individuals and nominal groups. The same is true when slight increases or decreases in performance can have dramatic consequences, for example because they make the difference between a near-accident and a nuclear GSA. However, when time is of the essence and/or small differences in performance do not have severe consequences, an elegant solution could be to mimic our nominal group condition. The statistical aggregation effect accounted for between 65% and 80% of the total performance gain we observed in real groups. In contrast, the time required to reap the benefits of statistical aggregation accounted for less than 20% of the additional time requirements of real groups. In other words, the first 20% of the additional time yielded at least 65% of the observed accuracy gains. The caveat here is that there may be cases in which averaging nominal group members' individual inputs does not improve accuracy. Averaging is a potent strategy when group members' individual errors are (largely) uncorrelated (Soll & Larrick, 2009). The more group members share certain biases, and the more their errors correlate, the less effective statistical aggregation will be. In those cases, nominal groups are unlikely to be an effective means to increase decision quality (Einhorn et al., 1977). Real groups, however, can still prove valuable in such situations. A recent study on quantitative judgments showed that even when group members' errors are highly correlated, real groups still benefit from synergy (Stern et al., 2017).

In this context, it is important to remember that the average model we used to compute the group potential, while constituting a challenging benchmark for the real groups, can theoretically be outperformed by more elaborate aggregation strategies. In particular, one can leverage the beneficial effects of statistical aggregation in nominal groups by differential weighting (just as differential weighting can be beneficial in real groups). For example, Mannes et al. (2014) found that small subsets of a crowd that were selected based on good prior performance were able to outperform the whole crowd in a wide range of situations. As Davis-Stober et al. (2014) show, it is a mathematical truth that there is always some weighted mean that performs better than the unweighted average of nominal group members' individual decisions. In the best case, differentially weighting nominal group members' individual decisions based on their members prior performance might allow nominal groups to match or even surpass the performance of real groups. However, our own data suggest that the advantages of differentially weighting nominal group members might vary with the task. Specifically, taking the initial best member of the nominal group instead of averaging all nominal group members' inputs yielded superior performance in Experiments 1 and 3, but inferior performance in Experiment 2 (see Supplementary Analysis S3 in the online supplemental materials).

Of course, one could employ more sophisticated weighting strategies than assigning all weight to the nominal groups' expert, but such strategies require sufficient information about nominal group members' past performance. For example, the model by Davis-Stober et al. (2014) requires accurate knowledge about nominal group members' individual expertise, their idiosyncratic biases, as well as the correlation of these biases. The more information about nominal group members' past performance is scarce or unreliable, the greater is the risk to misperceive nominal group members' relative expertise and, thus, to assign the wrong weights to their individual contributions. In those situations, aggregation strategies that are theoretically superior to averaging might actually backfire and yield inferior performance (Bednarik & Schultze, 2015; Dawes, 1979). Accordingly, we would advise to rely on differentially weighted nominal groups when managing dynamic system only if there is enough information to compute the optimal weights, and to rely on interacting groups otherwise.

Finally, we point to an important secondary benefit of working on dynamic system in real groups. Because of G-I transfer, participants who previously controlled the system as part of an interacting group were subsequently more adept at controlling the system than members of nominal groups or individuals were. Hence, if one of the primary goals is to train individuals to control a system successfully on their own, social interaction in a group seems to be much more effective than individual practice, and practitioners should consider this when deciding whether it is worth to have groups perform such a task.

Limitations and Directions for Future Research

The dynamic system simulations used in this laboratory study include characteristic features of technical, biological, and economic systems in real life, such as multiple opaque interconnected variables, nonlinear relationships between variables, and time delay of system reactions. Therefore, we are confident that our findings generalize to many real-world task contexts. However, we can think of a number of methodological limitations of our study that require consideration.

First, all experiments exclusively used three-person groups, and it remains to be tested whether or to what degree group size has an effect on the magnitude of the observed benefits of group work. Previous studies that assessed group performance under varying group sizes in dynamic control tasks (Wolfe & Chacko, 1983) or static problem solving tasks (Laughlin et al., 2006) showed that three-person groups outperformed dyads, but that more than three members did not add further performance improvements. Accordingly, it is plausible to assume that process gains of dyads will be smaller, and that G-I transfer in dyads might also be less pronounced due to the fact that less models for social learning are present in the group—but this is a question that has to be tested in further research.

Third, we used student samples that formed ad hoc groups. Whereas the average cognitive ability level and academic background is likely to be comparable with professionals who are entrusted with controlling dynamic systems, ad hoc groups' performance might not necessarily measure up to that of to established work groups that have developed and stayed together over longer periods. Members of ad hoc groups do not know each other personally and do not expect to collaborate in the future, which might influence their interaction behavior. Furthermore, established groups will develop a transactive memory system (Wegner, 1986), which should enable group members to track each other's individual capabilities and, as a consequence, to give greater weight to the contributions of more capable members. Thus, future research should systematically vary the duration of previous group interaction as well as transactive memory contents to test whether these factors foster differential weighting as a second form of group learning.

The third methodological limitation of our research is the self-paced input of control decisions and of corresponding system reactions in both scenarios used. Groups (and individuals) could discuss without time constraints before agreeing upon a control strategy and entering the input values to the system. As we have outlined, it took interacting groups significantly longer than individuals and nominal groups to accomplish the control test phase. However, many real-life contexts like air traffic control (cf. Gonzalez, Vanyukov, & Martin, 2005) or fighting large forest fires (cf. Omodei & Wearing, 1993, 1995) demand control judgments under time pressure. In many cases, a dynamic system does not wait for the decision-maker to alter the input variables. Hence, future studies should investigate whether similar manifestations of process gains and G-I transfer also occur when groups make control decisions under time restrictions and under a pacing that is determined by the system rather than by the decision-maker.

A fourth limitation of the current study is the range of complexity of the employed system scenarios. Although the inventory management task of Experiment 2 was more difficult than the cooling system in Experiments 1 and 3, both scenarios—if we take the whole range of dynamic system control tasks into account—are rather similar in terms of complexity. In comparison with early research on complex problem solving, the systems we used are on the lower end of the complexity continuum, because they involved only one input and one or two output variables. At the same time, time-delayed and nonlinear variable relations make our systems more complex than some systems used in contemporary research. As such, we cannot rule out that we hit a sweet spot of system complexity that allowed real groups to shine. Hence, an important question for future studies is whether our findings generalize to systems of different complexity. This includes more complex systems with multiple input variables and several—and potentially conflicting—outcome variables, like single firm marketing strategy simulations (Gray, 2012), corporation simulations with several business units (Fandt, Richardson, & Conner, 1990), political control scenarios (Dörner et al., 1983), or national economy simulations (Broadbent & Ashton, 1978). However, it also concerns dynamics systems with linear and immediate variable relations (e.g., Osman et al., 2015).

Related to the previous aspect, in the systems we used, input decisions were quantitative in nature. However, there are systems in which, at least, some decisions are categorical in nature. These decisions can work as on-off switches. For example, in a business setting, there might be a situation in which decision-makers can decide to implement (or revoke) certain incentive schemes for good performance, or they can switch between different marketing strategies to advertise their products. We consider it an important venue for future research to study group performance in such systems. A potential difficulty in doing so stems from categorical decisions differing from quantitative decisions in one important way: because the decision-makers must choose exactly one of several options, different aggregation rules are required to harness the wisdom of the crowds. A common method is the majority rule. Given that group members can choose the best option individually above chance level, the majority rule improves decision quality as group size increases. The benefits of majority voting have been shown in several contexts such as lie detection (Klein & Epley, 2015) or diagnosing cancer (Kurvers, Krause, Argenziano, Zalaudek, & Wolf, 2015), but also in dynamic decision making under uncertainty (Hastie & Kameda, 2005). The majority rule has certain similarities to the average model in quantitative decision-making. First, it requires no knowledge about group members' relative expertise. Second, it provides a challenging performance baseline because it can outperform not only individuals but also the best individual member of a nominal group (Hastie & Kameda, 2005; Kurvers et al., 2015). Finally, as with the average model, it is possible to outperform majority voting given additional information that interactive groups could acquire during discussion. For example, aggregation rules based on confidence as an indicator of expertise are particularly effective (Hertwig, 2012; Koriat, 2012). Because of these conceptual similarities between majority voting and averaging, we suggest using majority voting as the group potential in dynamic systems with categorical decisions.

Finally, the tasks we used arguably favored decision-makers who engaged in careful exploration and introduced relatively modest changes of the input variables. As such, they may have favored groups whose strategy consisted in choosing some sort of central tendency when making their joint decision. Such strategies are quite frequent in group decision-making to the point where some researchers consider them heuristics (Hertwig, 2012). However, some dynamic systems may require extreme decisions in order to ensure good system control performance. Accordingly, an important venue for future research would be to investigate group performance in dynamic systems that require extreme decisions and to test whether our findings generalize to such systems.

Conclusion

Our study provides the first unequivocal answer to the question of how well groups perform at complex system control tasks when compared with individuals. Interacting groups are, indeed, superior to individuals. This superiority goes beyond the simple aggregation of multiple individual judgments, that is, part of the performance advantage is the result of true synergy. Furthermore, we have shown that synergy is the result of individual capability gains in interacting groups that exceed mere practice effects or individual improvements due to aggregated reference information. Thus, it seems advisable to entrust groups with dynamic system control decisions to bolster performance outcomes and to improve the individual performance level of employees' dynamic decision-making.

Context

This work builds on our previous research on group judgment and decision-making. In this previous work, we investigated relatively simple tasks—such as estimating quantities or choosing between two or three options based on a given set of information—to investigate groups' ability to achieve synergy. Equipped with a relatively good understanding of the group processes that can, in principle, lead to synergy in group judgment and decision-making, we then felt that it was time to turn to more complex tasks such as dynamic system control. Although the search for synergy in dynamic system control is certainly methodologically challenging, it is also highly relevant because complex dynamic systems are ubiquitous in politics, health care, and business settings. Given that prior work comparing group and individual performance in such systems was scarce and that the few existing works had certain methodological shortcomings, our main motivation was to provide an unequivocal test of whether groups perform better than individuals when managing dynamic systems and, if so, whether this performance advantage is actually the result of synergy.

References

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight: The International Journal of Applied Forecasting, 5,* 3–15.

Badke-Schaub, P. (1993). *Gruppen und komplexe Probleme: Strategien von Kleingruppen bei der Bearbeitung einer simulierten AIDS-Ausbreitung* [Groups and complex problems: Strategies of small groups working with a simulated Aids epidemic]. Frankfurt am Main, Germany: Peter Lang.

Baumann, M. R., & Bonner, B. L. (2004). The effects of variability and expectations on utilization of member expertise and group performance. *Organizational Behavior and Human Decision Processes, 93,* 89–101. http://dx.doi.org/10.1016/j.obhdp.2003.12.004

Bednarik, P., & Schultze, T. (2015). The effectiveness of imperfect weighting in advice taking. *Judgment and Decision Making, 10,* 265–276.

Bonner, B. L., Sillito, S. D., & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior and Human Decision Processes, 103,* 121–133. http://dx.doi.org/10.1016/j.obhdp.2006.05.001

Borkenau, P., & Ostendorf, F. (1989). Untersuchungen zum Fünf-Faktoren-Modell der Persönlichkeit und seiner diagnostischen Erfassung [Investigations of the five-factor model of personality and its assessment]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 10,* 239–251.

Brehmer, B. (2005). Micro-worlds and the circular relation between people and their environment. *Theoretical Issues in Ergonomics Science, 6,* 73–93. http://dx.doi.org/10.1080/14639220512331311580

Broadbent, D. E., & Ashton, B. (1978). Human control of a simulated economic system. *Ergonomics, 21,* 1035–1043. http://dx.doi.org/10.1080/00140137808931810

Brodbeck, F. C., & Greitemeyer, T. (2000). A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes & Intergroup Relations, 3,* 159–182. http://dx.doi.org/10.1177/1368430200003002004

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision, 1,* 79–101. http://dx.doi.org/10.1037/dec0000004

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. http://dx.doi.org/10.1037/0003-066X.34.7.571

Dentler, P. (1977). *Zum Problemlöseverhalten von Gruppen in komplexen Problembereichen* [On problem-solving behavior of groups in complex problems] (Unpublished doctoral dissertation). University of Kiel, Germany.

Devine, D. J., Clayton, L. D., Philips, J. L., Dunford, B. B., & Melner, S. B. (1999). Teams in organizations: Prevalence, characteristics, and effectiveness. *Small Group Research, 30,* 678–711. http://dx.doi.org/10.1177/104649649903000602

Diehl, E., & Sterman, J. D. (1995). Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes, 62,* 198–215. http://dx.doi.org/10.1006/obhd.1995.1043

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8,* 1153. http://dx.doi.org/10.3389/fpsyg.2017.01153

Dörner, D., Kreuzig, H. W., Reiter, F., & Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität* [Lohhausen. On dealing with uncertainty and complexity]. Bern, Switzerland: Huber.

Dörner, D., & Reither, F. (1978). Über das Problemlösen in sehr komplexen Problembereichen [Problem solving in highly complex, real situations]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 25,* 527–551.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors, 4,* 59–74. http://dx.doi.org/10.1177/001872086200400201

Einhorn, H. J., Hogarth, R. N., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84,* 158–172. http://dx.doi.org/10.1037/0033-2909.84.1.158

Endres, J., & Putz-Osterloh, W. (1994). Komplexes Problemlösen in Kleingruppen: Effekte des Vorwissens, der Gruppenstruktur und der Gruppeninteraktion [Complex problem solving in small groups: Effects of preknowledge, group structure and group interaction]. *Zeitschrift für Sozialpsychologie, 25,* 54–70.

Fandt, P. M., Richardson, W. D., & Conner, H. M. (1990). The impact of goal setting on team simulation experience. *Simulation & Gaming, 21,* 411–422. http://dx.doi.org/10.1177/104687819002100405

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences of the United States of America, 108,* E625. http://dx.doi.org/10.1073/pnas.1109947108

Fredrickson, J. W., & Mitchell, T. R. (1984). Strategic decision processes: Comprehensiveness and performance in an industry with an unstable environment. *Academy of Management Journal, 27,* 399–423.

Funke, J. (1985). Steuerung dynamischer Systemedurch Aufbau und Anwendung subjektiver Kausal-modelle [Controlling dynamic systems throughacquisition and application of individual mentalmodels]. *Zeitschrift fur Psychologie mit Zeitschrift fur Angewandte Psychologie, 193,* 443–465.

Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. In R. J. Sternberg & P. A. Frensch (Eds.), *Complex problem solving: Principles and mechanisms* (pp. 185–222). Hillsdale, NJ: Erlbaum.

Funke, J. (1992). *Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung* [Knowledge about dynamic systems: Acquisition, representation, and use]. Berlin, Germany: Springer.

Funke, J. (2012). Complex problem solving. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 682–685). Heidelberg, Germany: Springer. http://dx.doi.org/10.1007/978-1-4419-1428-6_685

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121,* 149–167. http://dx.doi.org/10.1037/0033-2909.121.1.149

Gonzalez, C., Thomas, R. P., & Vanyukov, P. (2005). The relationships between cognitive ability and dynamic decision making. *Intelligence, 33,* 169–186. http://dx.doi.org/10.1016/j.intell.2004.10.002

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior, 21,* 273–286. http://dx.doi.org/10.1016/j.chb.2004.02.014

Gray, D. (2012). The influence of complexity and uncertainty on self-directed team learning. *International Journal of Learning and Change, 6,* 79–96. http://dx.doi.org/10.1504/IJLC.2012.045858

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36,* 189–213. http://dx.doi.org/10.1177/0146621612439620

Greiner, B. (2004). An online recruitment system for economic experiments. In K. Kremer & V. Macho (Eds.), *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63* (pp. 79–93). Goettingen, Germany: Ges. fuer Wiss. Datenverarbeitung.

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, NY: Academic Press. http://dx.doi.org/10.1016/S0065-2601(08)60248-8

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review, 112,* 494–508. http://dx.doi.org/10.1037/0033-295X.112.2.494

Hertwig, R. (2012). Psychology. Tapping into the wisdom of the crowd—With confidence. *Science, 336,* 303–304. http://dx.doi.org/10.1126/science.1221403

Hofinger, G., Rek, U., & Strohschneider, S. (2006). Menschengemachte Umweltkatastrophen—Psychologische Hintergründe am Beispiel von Tschernobyl [Psychological factors in man-made environmental catastrophes: The Tschernobyl example]. *Umweltpsychologie, 10,* 26–45.

Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology, 10,* 371–384. http://dx.doi.org/10.1016/0022-1031(74)90033-X

Keck, S., & Tang, W. (2019). Elaborating or aggregating? The joint effects of decision-making structure and systematic errors. *Academy of Management Proceedings, 2019,* 12856. http://dx.doi.org/10.5465/AMBPP.2019.299

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55,* 623–655. http://dx.doi.org/10.1146/annurev.psych.55.090902.142009

Kirlik, A., Plamondon, B. D., Lytton, L., Jagacinski, R. J., & Miller, R. A. (1993). Supervisory control in a dynamic and uncertain environment: Laboratory task and crew performance. *IEEE Transactions on Systems, Man, and Cybernetics, 23,* 1130–1138. http://dx.doi.org/10.1109/21.247893

Klein, N., & Epley, N. (2015). Group discussion improves lie detection. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 7460–7465. http://dx.doi.org/10.1073/pnas.1504048112

Kluge, A. (2014). *The acquisition of knowledge and skills for taskwork and teamwork to control complex technical systems: A cognitive and macroergonomics perspective*. Dordrecht, the Netherlands: Springer. http://dx.doi.org/10.1007/978-94-007-5049-4

Köller, O., Dauenheimer, D. G., & Strauß, B. (1993). Unterschiede zwischen Einzelpersonen und Dyaden beim Lösen komplexer Probleme in Abhängigkeit von der Ausgangsfähigkeit [Differences between individuals and two-person groups in solving complex problems in relation to initial ability level]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 40,* 194–221.

Koriat, A. (2012). When are two heads better than one and why? *Science, 336,* 360–362. http://dx.doi.org/10.1126/science.1216549

Kurvers, R. H., Krause, J., Argenziano, G., Zalaudek, I., & Wolf, M. (2015). Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *Journal of the American Medical Association Dermatology, 151,* 1346–1353. http://dx.doi.org/10.1001/jamadermatol.2015.3149

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. http://dx.doi.org/10.1037/0022-3514.37.6.822

Laughlin, P. R., Carey, H. R., & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations, 11,* 319–330. http://dx.doi.org/10.1177/1368430208090645

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology, 90,* 644–651. http://dx.doi.org/10.1037/0022-3514.90.4.644

Laughlin, P. R., Zander, M. L., Knievel, E. M., & Tan, T. K. (2003). Groups perform better than the best individuals on letters-to-numbers problems: Informative equations and effective strategies. *Journal of Personality and Social Psychology, 85,* 684–694. http://dx.doi.org/10.1037/0022-3514.85.4.684

Leutner, D. (1988). Computersimulierte dynamische Systeme: Wissenserwerb unter verschiedenen Lehrmethoden und Sozialformen des Unterrichts [Computer simulated dynamic systems: Knowledge acquisition under different teaching methods and classroom formats]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 20,* 338–355.

Lippold, M. A., Schulz-Hardt, S., & Schultze, T. (in press). G-I transfer in multi-cue judgment tasks: Discussion improves group members' knowledge about target relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Maciejovsky, B., & Budescu, D. V. (2007). Collective induction without cooperation? Learning and knowledge transfer in cooperative groups and competitive auctions. *Journal of Personality and Social Psychology, 92,* 854–870. http://dx.doi.org/10.1037/0022-3514.92.5.854

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107,* 276–299. http://dx.doi.org/10.1037/a0036677

Minson, J. A., Mueller, J. S., & Larrick, R. P., & the Undermines the Accuracy of Collaborative Judgments. (2018). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science, 64,* 4177–4192. http://dx.doi.org/10.1287/mnsc.2017.2823

Omodei, M. M., & Wearing, A. J. (1993). *Fire chief user manual*. Melbourne, Australia: Department of Psychology, University of Melbourne.

Omodei, M. M., & Wearing, A. J. (1995). The Fire chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision making behavior. *Behavior Research Methods, Instruments, & Computers, 27,* 303–316. http://dx.doi.org/10.3758/BF03200423

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin, 136,* 65–86. http://dx.doi.org/10.1037/a0017815

Osman, M., Glass, B. D., & Hola, Z. (2015). Approaches to learning to control dynamic uncertainty. *Systems, 3,* 211–236. http://dx.doi.org/10.3390/systems3040211

Osman, M., Glass, B. D., Hola, Z., & Stollewerk, S. (2017). Reward and feedback in the control over dynamic events. *Psychology, 8,* 1063–1089. http://dx.doi.org/10.4236/psych.2017.87070

Osman, M., & Palencia, D. O. V. (2019). The future of problem solving research is not complexity, but dynamic uncertainty. *Journal of Dynamic Decision Making*. Advance online publication. http://dx.doi.org/10.11588/jddm.2019.1.69300

Paich, M., & Sterman, J. D. (1993). Boom, bust and failures to learn in experimental markets. *Management Science, 39,* 1439–1458. http://dx.doi.org/10.1287/mnsc.39.12.1439

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [The relation between test intelligence and problem solving success]. *Zeitschrift fur Psychologie mit Zeitschrift fur Angewandte Psychologie, 189,* 79–100.

Reason, J. (1987). The Chernobyl errors. *Bulletin of the British Psychological Society, 40,* 201–206.

Reichert, U., & Dörner, D. (1988). Heurismen beim Umgang mit einem "einfachen" dynamischen System [Heuristics of coping with a "simple" dynamic system]. *Sprache und Kognition, 7,* 12–24.

Rouwette, E. A. J. A., Größler, A., & Vennix, J. A. M. (2004). Exploring influencing factors on rationality: A literature review of dynamic decision-making studies in system dynamics. *Systems Research and Behavioral Science, 21,* 351–370. http://dx.doi.org/10.1002/sres.647

Schlemmer, A. (2009). Assembly lab for experiments (Versions 2009 to 2011) [Computer Software]. Göttingen, Germany: GNU General Public License Version 3.

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative. *Organizational Behavior and Human Decision Processes, 118,* 24–36. http://dx.doi.org/10.1016/j.obhdp.2011.12.006

Schultze, T., & Schulz-Hardt, S. (2020). *Synergy in dynamic system control*. Retrieved from osf.io/e3uky

Schulz-Hardt, S., & Mojzisch, A. (2012). How to achieve synergy in group decision-making: Lessons to be learned from the hidden profile paradigm. *European Review of Social Psychology, 23,* 305–343. http://dx.doi.org/10.1080/10463283.2012.744440

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43,* 1–28. http://dx.doi.org/10.1016/0749-5978(89)90055-1

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 780–805. http://dx.doi.org/10.1037/a0015145

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science, 11,* 273–283. http://dx.doi.org/10.1002/bs.3830110404

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

Sterman, J. D. (1989a). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes, 43,* 301–335. http://dx.doi.org/10.1016/0749-5978(89)90041-1

Sterman, J. D. (1989b). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science, 35,* 321–339. http://dx.doi.org/10.1287/mnsc.35.3.321

Stern, A., Schultze, T., & Schulz-Hardt, S. (2017). How much group is necessary? Group-to-individual transfer in estimation tasks. *Collabra: Psychology, 3,* 16. http://dx.doi.org/10.1525/collabra.95

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York, NY: Doubleday.

van Ginkel, W. P., & van Knippenberg, D. (2008). Group information elaboration and group decision making: The role of shared task representations. *Organizational Behavior and Human Decision Processes, 105,* 82–97. http://dx.doi.org/10.1016/j.obhdp.2007.08.005

Wegner, D. M. (1986). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). New York, NY: Springer.

Wolfe, J., & Chacko, T. I. (1983). Team-size effects on business game performance and decision making behaviors. *Decision Sciences, 14,* 121–133. http://dx.doi.org/10.1111/j.1540-5915.1983.tb00173.x

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science, 13,* 75–78. http://dx.doi.org/10.1111/j.0963-7214.2004.00278.x

Zajonc, R. B. (1965). The requirements and design of a standard group task. *Journal of Experimental Social Psychology, 1,* 71–88. http://dx.doi.org/10.1016/0022-1031(65)90037-5

Ziegler, R., Diehl, M., & Zijlstra, G. (2000). Idea production in nominal and virtual groups: Does computer-mediated communication improve group brainstorming? *Group Processes & Intergroup Relations, 3,* 141–158. http://dx.doi.org/10.1177/1368430200032003

(*Appendices follow*)

Appendix A

Underlying System Equations of the Cooling System Scenario in Experiments 1 and 3

We adapted the dynamic system control tasks used in Experiments 1 and 3 from Reichert and Dörner (1988). We solely changed the semantic context from a cooling system in a grocery store to a cooling system in a nuclear power plant, but left the specification of system otherwise unchanged. The system consists of two formulae linking participants' inputs to the system's internal variables as well as the system output.

$$temp_t = temp_{t-1} + (external - temp_{t-1}) \times insulation - control_{t-1} \tag{1}$$

$$control_t = (temp_{t-v} - input_t) \times inertia \tag{2}$$

In Equation (1), $temp_t$ is the system temperature at time t or, more specifically, a linear transformation of this temperature (to compute the temperature in °C, one needs to subtract 30 and then divide the value by 7.5). System temperature at time t is the sum of three components. The first is the system's temperature in the previous round, $temp_{t\,-\,1}$. Second, this old system temperature drifts toward the external temperature *external*, which is a constant set to 170 (or 18.66°C). The greater the difference between the systems previous temperature and the external temperature, the greater is the drift, but this drift is slowed down by the system's *insulation*, which is another constant set to a value of 0.10. Finally, the system temperature depends on $control_{t\,-\,1}$, which is an internal variable that depends on participants' inputs.

Equation (2) states how the internal variable *control* depends on participants' control decisions. $Input_t$ captures the current value of the input variable that participants can change. Possible values range from 0 to 200 with a starting value of 100. Finally, *inertia* is another constant representing the systems' inertia when reacting to participants' inputs. Inertia is fixed at a level of 0.30. *Control* also depends on previous levels of the system temperature *temp*. Here, v is a time delay parameter specifying the time lag for the influence of previous system temperatures on *control*. In this system, v is set to 3, which means that the temperature at any given point in time t depends on the temperature at t_{-1} (because of Equation 1), the temperature at t_{-4} (because of inserting Equation 2 into Equation 1), and the system input at t_{-1}. This has two consequences. First, it means that at times $t = 0$ (i.e., when participants start to work) till $t = 3$ *control* depends, in part, on system states prior to participants' first decision. These prior values of $temp_{-3} = 17$, $temp_{-2} = 62.3$, $temp_{-1} = 103.07$, and $temp_0 = 139.79$ create an internal temperature oscillation and, thus, a nonlinear system. Second, the system state will be identical for all participants at the end of the first trial because the effects of the first control decision will only come to bear one turn later.

Appendix B

Underlying System Equations of the Stock Management Scenario in Experiment 2

We adapted the dynamic system of Experiment 2 from Diehl and Sterman (1995). The underlying system formulae are as follows:

$$I_t = I_{t-1} + P_t - S_t \tag{3}$$

Equation (3) simply states that I_t, the company's inventory at time t, equals the previous inventory plus current production P_t minus current sales S_t.

$$P_t = P^*_{t-v} \tag{4}$$

$$P^*_t = P^*_{t-1} + \Delta_t \tag{5}$$

Equations (4) and (5) are concerned with the company's production. As Equation (4) shows, current production P_t equals the number of previously commissioned units P^*_{t-v}, where v represents the time it takes to produce the commissioned units (here, it takes the value 2). Finally, Equation (5) specifies that P^*_t, the number of units commissioned at time t, depend on the number of units commissioned in the previous round, P^*_{t-1}, and participants' decision about how to alter last round's number of commissioned units, represented by Δ_t. In other words, participant can increase the company's production rate by entering positive values of Δ and decreased it by entering negative values of Δ. Due to a natural lower limit of zero commissioned units, the computer automatically corrected values that would have resulted in negative values of P^*_t.

$$S_t = X_t + \gamma P_t \tag{6}$$

(*Appendices continue*)

The total amount of Sales are the sum of two components as can be seen from Equation (6). Exogenous Sales X_t correspond to the proportion of costumer demand that is not influenced by the production rate of the company. The exogenous component follows a random walk throughout the 40 trials of a test phase, which was generated once and then held constant for all phases and participants. This random walk started at a level of 400 units, and its round-by-round change was the result of draws from a uniform distribution of integers ranging from −15 to 15. The second component represents the endogenous sales. Here, parameter γ determines to what extent the company's production rate leads to increases (or decreased) demand. We set γ to a level of 0.30, leading to a positive feedback effect: increasing production lead to an increase in endogenous sales. This positive feedback effect corresponds to realistic market reactions known as Keynesian multiplier (cf. Diehl & Sterman, 1995, p. 201).

$$C_t = a \cdot I_t^2 + b \cdot \Delta_t^2 \tag{7}$$

Finally, Equation (7) determines the systems' central outcomes variable C_t, that is, the total costs incurred at time t. The first cost factor are inventory costs defined as the square of the respective turn's inventory, I_t^2. Positive values of I_t result in storage cost such as rent or wages, whereas negative values of I_t indicates that the company could not meet the demand on time, resulting in, for example, disappointed customers, image loss, and an increased customer churn rate. The second cost factor, Δ_t^2, contains all the costs associated with changing the company's production rate, such as hiring and training costs or severance payments for laid-off workers. Parameters a and b are simple scaling parameters set to values of 1 and 2, respectively.

Received October 18, 2018
Revision received July 12, 2020
Accepted July 26, 2020 ■