





## RESEARCH ARTICLE

WILEY

# Collaboration during the diagnostic decision-making process: When does it help?

Juliane E. Kämmer<sup>1,2</sup>  | Karin Ernst<sup>1</sup> | Kim Grab<sup>1</sup> | Stefan K. Schaubert<sup>3</sup>  |  
Stefanie C. Hautz<sup>1</sup>  | Dorothea Penders<sup>4,5</sup> | Wolf E. Hautz<sup>1</sup> 

<sup>1</sup>Department of Emergency Medicine, Inselspital, University Hospital Bern, University of Bern, Bern, Switzerland

<sup>2</sup>Department of Social and Communication Psychology, Institute for Psychology, University of Göttingen, Göttingen, Germany

<sup>3</sup>Centre for Health Sciences Education and Centre for Educational Measurement, University of Oslo, Oslo, Norway

<sup>4</sup>Learning Center, Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

<sup>5</sup>Department of Anesthesiology and Intensive Care Medicine, Charité-Universitätsmedizin Berlin, Berlin, Germany

## Correspondence

Juliane E. Kämmer, Department of Emergency Medicine, Inselspital, University Hospital Bern, Freiburgstrasse 16C, 3010 Bern, Switzerland. Email: [juliane.kaemmer@unibe.ch](mailto:juliane.kaemmer@unibe.ch)

## Funding information

Horizon 2020 Framework Programme, Grant/Award Numbers: 101021775, 894536

## Abstract

When making complex decisions, such as a medical diagnosis, decision makers typically gather, analyze, and synthesize (integrate) information. In a previous study, we showed that delegating such complex decisions to collaborating pairs increases decision quality substantially compared to that of individuals, without requiring different information gathering. Given the higher costs associated with teamwork, however, it is of great practical interest to understand when in the process the performance benefits of teams may arise, so that particular subtasks can be delegated to teams when most appropriate. We thus conducted an experimental study in which fourth-year medical students ( $n = 109$ ) worked either in pairs or alone on two separate subtasks of the diagnostic process: (1) analyzing diagnostic test results (e.g., X-rays) and (2) integrating previously interpreted test results into diagnoses. Linear mixed-effects models revealed a small benefit of collaborating pairs over individuals in both subtasks. We conclude that collaborating with a peer may pay off both when analyzing information *and* when integrating it into a diagnosis as it provides the opportunity to correct each other's errors and to make use of a greater knowledge base. These findings encourage the strategic use of collaboration with a colleague when making complex decisions. Further research into the underlying processes is needed.

## KEYWORDS

collaboration, complex problem solving, decision-making, information processing, medical diagnosis, teamwork

## 1 | INTRODUCTION

Diagnostic decisions are ubiquitous in private and professional life. Individuals, teams, and whole organizations are confronted with diagnostic problems every day, sometimes involving life and death. Think of problems in the fields of health care, business, justice, education, or politics, such as the following: What diagnosis does a patient have? Which candidate is the most suitable for a given job? What are the

prospects of a new venture? What sentence is appropriate for someone convicted of a crime? Most of these real-world problems require decisions under uncertainty, such that decisions have to be based on incomplete information or made under time pressure. Also, in most of these real-world settings, wrong decisions may have severe consequences, for example, medical diagnostic errors (i.e., wrong, delayed, or missed diagnoses) may lead to increased length of hospital stay and mortality (Hautz et al., 2019), thus posing a serious threat to patient

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Journal of Behavioral Decision Making* published by John Wiley & Sons Ltd.

safety and a burden to the health-care system (Berner & Graber, 2008; Singh et al., 2014; Zwaan et al., 2010). To better deal with the uncertainty and to take advantage of the increased specialization of experts, modern organizations rely on teams (Deloitte Insights, 2019; Edmondson, 2012; Salas, 2008). At universities, for instance, appointment committees collectively explore the suitability of candidates for a professorship, and in hospitals, usually teams of several health-care professionals engage in deciding on a diagnosis or treatment (Committee on Diagnostic Error in Health Care et al., 2015; Graber et al., 2017).

When and why are teams superior to individuals? These questions have received substantial attention in recent decades, particularly in the context of teamwork becoming the prevailing mode of work (Deloitte Insights, 2019). Extensive empirical and theoretical research has shed light on the superiority of teams over individuals and has highlighted their resources and also their pitfalls (Esser, 1998; Kerr & Tindale, 2004; Mathieu et al., 2017). For example, on the positive side, teams may outperform individuals because they have a greater capacity to attend to information, a larger joint memory for storing information, and a larger knowledge base (Hinsz et al., 1997; Vollrath et al., 1989). On the downside, team interactions may lead to social loafing (Karau & Williams, 1993), biased information search (Schulz-Hardt et al., 2000, 2002), or even groupthink (Turner & Pratkanis, 1998).

Research has suggested that the task has a fundamental impact on (team) performance (Steiner, 1972). In the realm of diagnostic tasks, for example, studies have demonstrated that pairs and larger collectives exhibit superior performance (Hautz et al., 2015; Kämmer et al., 2017). However, complex diagnostic decision-making entails multiple sequential and/or reiterative subtasks, beginning with gathering available information, proceeding to its analysis and interpretation, and culminating in integrating available evidence into a final diagnosis that serves as a basis for action (Committee on Diagnostic Error in Health Care et al., 2015). These subtasks likely have different requirements, which may lead to relative advantages of teams over individuals in some subtasks but not in others. Yet, our current understanding of the functioning and potential benefits of teams during those subtasks remains limited, primarily due to studies failing to differentiate between subtasks. However, it is of great theoretical and practical interest to determine which subtask or tasks within the diagnostic process yield the most significant advantages through collaboration.

First, from a theoretical standpoint and as highlighted by Larson (2010), a comprehensive understanding of overall performance in such complex tasks necessitates consideration of their underlying subtask structure. By acknowledging the subtask structure of complex tasks, one can gain a more nuanced understanding of the intricacies of decision-making processes and how task requirements interact with group resources and limitations. This analysis can reveal the extent to which mechanisms observed in groups, such as improved joint memory (Hinsz, 1990), contribute to subtask and overall performance.

Second, from a practical perspective, it is crucial to identify when errors and benefits arise in team settings compared to individual

settings. This knowledge can serve as an empirical basis for making strategic decisions on task delegation to teams, especially when faced with limited resources or time constraints. Additionally, this understanding is valuable for developing and refining interventions aimed at enhancing diagnostic decision-making. Practical applications, such as in the context of hospitals, where true teamwork time is often scarce and fragmented collaboration due to organizational constraints occurs (Olson et al., 2020), underscore the relevance of this knowledge. For instance, in teaching hospitals, junior physicians typically gather information and interpret it individually before collaborating with senior physicians to integrate information into a diagnosis and recommendation for further treatment. Similarly, in appointment committees, members usually review applicants' documents independently before convening to collectively assess the information and make a decision.

To find where in the diagnostic process performance benefits of teams occur, we conducted an experimental study in which we assessed the performance of individuals and teams in different decision-making subtasks. We focused on ill-defined tasks to which decision makers brought some prior knowledge and expertise—features that are typical of many real-world decisions.

In particular, we studied the diagnostic decision process in teams versus individuals with emergency medicine as our exemplary decision-making environment. Although emergency medicine may appear to be a unique and highly specific environment, we would argue that it shares many features with other high-risk environments (see also Hagemann et al., 2011). In fact, many other health-care settings (e.g., the operating theater) and other domains such as the business sector, the military, aerospace, and disaster relief organizations (DeChurch et al., 2011) can similarly be characterized as collaborative, rapidly evolving, uncertain environments with critical consequences of performance failures and great potential for performance gains. Moreover, making a medical diagnosis shares a number of features with other uncertain or ill-defined task environments, including that

1. access to data is heterogeneous (e.g., some data are readily available, whereas some data entail waiting time/costs/potential harm if testing is invasive; data need to be actively gathered from different sources);
2. there is a (sometimes unknown) probabilistic relationship between symptoms/diagnostic data (cues) and the correct diagnosis (criterion), and complex interdependencies between symptoms and diagnoses are possible;
3. the number of diagnoses to select from is uncertain and potentially very large; and
4. context factors such as time pressure, stress, noise, and multiple decisions at the same time, among others, accompany the decision process.

Given the relevance of these attributes for a range of organizations and settings, studying collaborative diagnostic decision-making in an emergency room setting has the potential to produce beneficial lessons for a broad range of domains. In the following, we provide the

relevant theoretical background for our study, drawing on decision psychology and clinical reasoning theories as well as small-group research.

## 1.1 | Task analysis of diagnosis decision-making

To make predictions about where in the process performance benefits of teams may arise, we performed a subtask analysis where we

mapped relevant group resources and limitations, which were identified by past work, to the subtask-specific requirements (for an overview see Table 1). Taking into account prominent cognitive psychological models of individual information processing (e.g., Gigerenzer & Gaissmaier, 2011; Massaro & Cowan, 1993), models of collective information processing (De Dreu et al., 2008; Hinsz et al., 1997; Propp, 1999), and models of clinical reasoning (Committee on Diagnostic Error in Health Care et al., 2015; Kiesewetter et al., 2017), we adopted a simple model of information

**TABLE 1** Task analysis and related group resources as well as challenges.

Task requirements/cognitive demands (in medical diagnostics)			Group resources/advantages of collaboration	Impaired processes due to collaboration
Information gathering	Information analysis and interpretation	Information integration		
Avoiding missing important pieces of information Directing attention and perception to relevant pieces of information and parts thereof (e.g., areas in visual material)			<ul style="list-style-type: none"> <li>Greater capacity to attend to information (Hinsz et al., 1997; Vollrath et al., 1989)</li> </ul>	<ul style="list-style-type: none"> <li>Biased information search due to motivational/preference-driven processing (Schulz-Hardt et al., 2000, 2002)</li> <li>Distraction and production blocking due to presence of others, social motives (De Dreu et al., 2008; Diehl &amp; Stroebe, 1987)</li> </ul>
Retrieving and applying factual (e.g., biomedical) knowledge/skills to decide where to search, what to search for, when to stop searching	Retrieving and applying factual (e.g., biomedical) knowledge skills systematically to material at hand and retrieving exemplars from memory		<ul style="list-style-type: none"> <li>Larger joint knowledge base (Hinsz et al., 1997; Vollrath et al., 1989)</li> <li>Mutual error correction and stimulation of knowledge retrieval possible (Collins &amp; Guetzkow, 1964; Hinsz, 1990; Vollrath et al., 1989)</li> <li>Enhanced information processing via the impact of the presence of others on accountability (Lerner &amp; Tetlock, 1999)</li> <li>Activation and restructuring of prior knowledge through verbalization (Olson et al., 2020)</li> </ul>	<ul style="list-style-type: none"> <li>Impaired knowledge retrieval, memory errors (Rajaram &amp; Pereira-Pasarin, 2010)</li> <li>Decreased information processing as a result of social loafing (Karau &amp; Williams, 1993) and diffusion of responsibility processes (Petty et al., 1980)</li> </ul>
	Recognizing and identifying patterns (e.g., “illness scripts”)		<ul style="list-style-type: none"> <li>Groups are capable of recognizing the truth if it is proposed by at least one member; i.e., “truth wins” (Larson, 2010; Laughlin et al., 1991)</li> <li>Groups are capable of recognizing emergent solutions including the identification of patterns not recognized by an individual (Laughlin et al., 1991, 2006)</li> </ul>	<ul style="list-style-type: none"> <li>Conflicts if different mental models present (Mathieu et al., 2000)</li> </ul>

(Continues)

TABLE 1 (Continued)

Task requirements/cognitive demands (in medical diagnostics)				
Information gathering	Information analysis and interpretation	Information integration	Group resources/advantages of collaboration	Impaired processes due to collaboration
Storing acquired information in short-term memory to guide further search or to decide when sufficient information has been acquired		Storing information in short-term memory until making a diagnosis Recalling given information	<ul style="list-style-type: none"><li>• Larger joint memory to store information (Hinsz et al., 1997)</li><li>• Groups are better at recalling given information (Vollrath et al., 1989)</li><li>• Groups are capable of processing more information than individuals (Laughlin et al., 1991)</li><li>• Groups are more likely than individuals to correct errors of memory (Hinsz, 1990; Vollrath et al., 1989)</li></ul>	<ul style="list-style-type: none"><li>• Difficulty with sharing privately held/distributed information (Boos et al., 2013; Larson et al., 1994, 1996, 1998)</li><li>• Collaborative inhibition due to retrieval interference or social loafing (Weldon et al., 2000)</li></ul>
		Weighting and using multiple pieces of information to test diagnostic hypotheses	<ul style="list-style-type: none"><li>• Larger joint knowledge base (Hinsz et al., 1997; Vollrath et al., 1989)</li><li>• More consistently applied rules (Chalos &amp; Pickard, 1985)</li><li>• Groups are more likely to recognize valid information (Lorge &amp; Solomon, 1955)</li><li>• Groups are more likely to reject erroneous information (Hirokawa &amp; Pace, 1983)</li></ul>	<ul style="list-style-type: none"><li>• Biased weighting due to striving for unanimity/conformity (Turner &amp; Pratkanis, 1998)</li></ul>

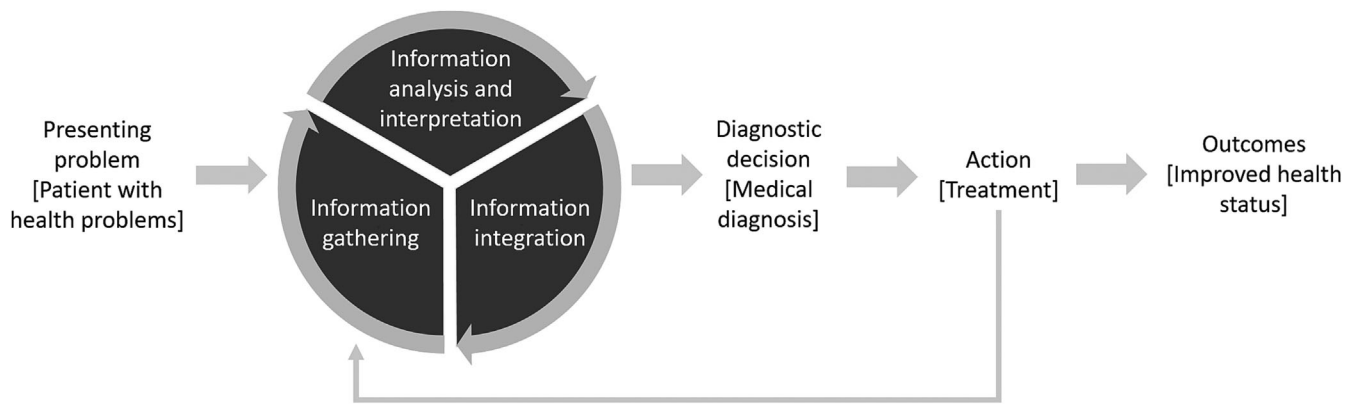


FIGURE 1 Simple stage model of the diagnostic decision-making process (medical example in square brackets). Adapted from Committee on Diagnostic Error in Health Care et al. (2015).

processing during the diagnostic process (Figure 1). This model encompasses the following subtasks: information gathering, information analysis and interpretation, and information integration. Following the common conceptualization of groups as information

processors (Hinsz et al., 1997), we assumed that teams go through the same subtasks as individuals do. By drawing the process as a circle (Figure 1), we acknowledge that these subtasks may be taken reiteratively before a final decision can be made.

First, during the information-gathering subtask, the diagnosticians search for information in the environment, for example, a physician performs a clinical interview with a patient, conducts a physical exam, or administers diagnostic tests to gain information about the patient's state. This subtask requires diagnosticians to direct their attention and perception to relevant pieces of information, to avoid missing relevant information, and to retrieve and apply knowledge and skills to decide where to search, what to search for, and when to stop searching. With regard to these task requirements, teams may outperform individuals because of their greater capacity to attend to information, their larger joint knowledge base, enhanced information processing, and the possibility of correcting each other's errors (Hinsz, 1990; Hinsz et al., 1997; Vollrath et al., 1989). Access to more information, in turn, may help teams entertain and test more hypotheses and thus prevent premature closure, that is, limiting the hypothesis space too early (Krupat et al., 2017). At the same time, teams may be impaired by process losses such as distraction and production blocking, or biased information search due to motivational processing (Schulz-Hardt et al., 2000, 2002).

Second, during the information-analysis subtask, the acquired information is (consciously) perceived, interpreted, and evaluated using available knowledge. For example, relying on their knowledge and expertise, a physician might evaluate an electrocardiogram (ECG) as abnormal and interpret it as showing signs of a certain disease. This subtask requires diagnosticians to direct attention to relevant parts or pieces of information and to retrieve and apply knowledge and skills to make sense of the given information. It also requires diagnosticians to identify patterns in the (visual) material (e.g., an X-ray). Here, teams may benefit particularly from their team members' ability to recognize/adopt the correct solution if it is proposed by at least one member (i.e., "truth wins"), and they are capable of recognizing emergent solutions (Laughlin et al., 1991). At the same time, they may be impaired by conflicts that arise if different mental models are present (Mathieu et al., 2000).

Third, during the information-integration subtask, the available evidence is weighted and used (i.e., integrated) to make a preliminary or final decision. For example, the physician comes up with a working diagnosis that may explain the patient's symptoms, taking into account the collected, interpreted evidence. This subtask requires diagnosticians to retrieve the acquired and interpreted evidence from short-term memory to make a diagnosis. This subtask also requires them to apply stored knowledge to appropriately weight and use the available pieces of information in order to verify or falsify certain diagnostic hypotheses, and to identify patterns in the information set (e.g., "illness scripts," which are general representations of an illness in the diagnostician's mind). For this subtask, groups may particularly benefit from their larger joint knowledge base (Hinsz et al., 1997; Vollrath et al., 1989), their better ability to recall information (Vollrath et al., 1989), and the higher likelihood of recognizing valid versus erroneous information (Lorge & Solomon, 1955). At the same time, they may be impaired by difficulties with sharing privately held information (Boos et al., 2013; Larson et al., 1994, 1996, 1998) and biased weighting when striving for unanimity (Turner & Pratkanis, 1998).

In summary, this task analysis highlights that there are differences and also commonalities in task requirements among the three subtasks. Consequently, similar group resources and challenges will likely impact performance in individual subtasks, rendering it challenging to predict the comparative performance benefits of teams in specific subtasks. Previous studies have primarily focused on either the complete diagnostic task or, at most, the information-gathering subtask (Hinsz, 1990; Schulz-Hardt et al., 2000), leaving performance differences between individuals and teams in the information-analysis and information-integration subtasks unexplored. Therefore, our objective was to investigate observable performance differences before delving into further research to examine the underlying mechanisms.

## 1.2 | Information gathering in pairs versus individuals

In a previous study, we aimed to establish whether making a diagnosis as a team enhances diagnostic accuracy compared to making it alone and whether differences in the information-gathering subtask account for potential performance differences (Hautz et al., 2015). We studied pairs of peers, a constellation that is common in many workplace settings. Specifically, we conducted an experimental lab study involving  $N = 88$  fourth-year medical students who had to make diagnoses for six simulated patients (Kunina-Habenicht et al., 2015) in one of two conditions, alone or as a partner in a collaborating pair. The experimental setup mirrored the real-world diagnostic decision-making process with its self-determined information gathering, analysis, and integration subtasks. Participants (either alone or in pairs, depending on the experimental condition) first read a short vignette containing basic information about a patient entering the emergency room with shortness of breath. Their task was then to diagnose the patient as quickly as reasonably possible using up to 30 pieces of information they were free to obtain in any order or ignore. Information was provided in text form (e.g., pulse rate), as an image (e.g., chest X-ray), or in audio form (e.g., heart sound) and thus had to be analyzed first.

As reported elsewhere (Hautz et al., 2015), we found that collaborating pairs ( $M = 67.78\%$ ) outperformed individuals ( $M = 50.00\%$ ), indicating that collaboration during the decision process had a large positive effect on diagnostic accuracy ( $d = 0.78$ ). Comparing collaborating pairs with nominal pairs ( $M = 56.73\%$ ) further indicated that we had observed a synergistic effect, which is defined as a performance gain due to group interaction (Larson, 2010, p. 4). This benefit, however, was not associated with enhanced information gathering, as collaborating pairs acquired neither more nor more relevant diagnostic tests. Thus, performance benefits may have arisen because of better information analysis or integration, or a combination of both. The experimental setup of this study, however, did not allow us to attribute the origin(s) of performance benefits to either of the subtasks. This is where the current study started.

### 1.3 | Overview of the study

The current study aimed to shed light on the question of when in the diagnostic decision process collaborating peers outperform individual decision makers. Specifically, we focused on the information-analysis and information-integration subtasks. To identify positive and negative effects of social interaction on performance, we compared not only real (i.e., interacting) pairs with individuals but also real pairs with nominal pairs (Larson, 2010). Drawing from the same population as in our prior work (Hautz et al., 2015), we studied advanced medical students who brought medical expertise to the experimental tasks. To gain insights into the underlying mechanisms, we also analyzed the complexity of reasoning.

## 2 | MATERIALS AND METHODS

### 2.1 | Study design

The study was conducted in two waves between May 2019 and January 2021 at Charité Medical School in Berlin, where fourth-year medical students were asked via mailing lists to participate in exchange for financial remuneration (€35, \$42 USD at that time). Although data collection in 2019 was organized as on-site lab sessions, data collection in December 2020/January 2021 had to be organized as online sessions via MS Teams, owing to contact restrictions during the COVID-19 pandemic. The study had a  $2 \times 2$  factorial design, in which decision source (individual vs. pair) was a between-subjects factor and task (analysis vs. integration) was a within-subject factor. We report all measures, manipulations, and exclusions in this study.

### 2.2 | Participants

The participants were 109 students from Charité Medical School in Berlin, with  $n = 39$  randomized into the individual condition ( $M_{\text{age}} = 22.77$  years,  $SD = 2.25$ , 51% female) and  $n = 70$  into the pair condition ( $M_{\text{age}} = 23.46$  years,  $SD = 3.1$ , 67% female; see Table 2 for details).

## 3 | PROCEDURE

Prior to the random assignment to work individually or in pairs (of the same gender), students gave their written informed consent. They then answered demographic questions and filled out a 22-item multiple-choice test that assessed their medical knowledge about shortness of breath (knowledge pretest). Afterward, they received a software demonstration and thorough instructions for the two tasks that followed. The participants then worked—either alone or in pairs, according to the condition they were in—on the integration task and then on the analysis task, with items presented in random order per task. There were no time restrictions. Pairs were instructed to arrive at answers together for all items.

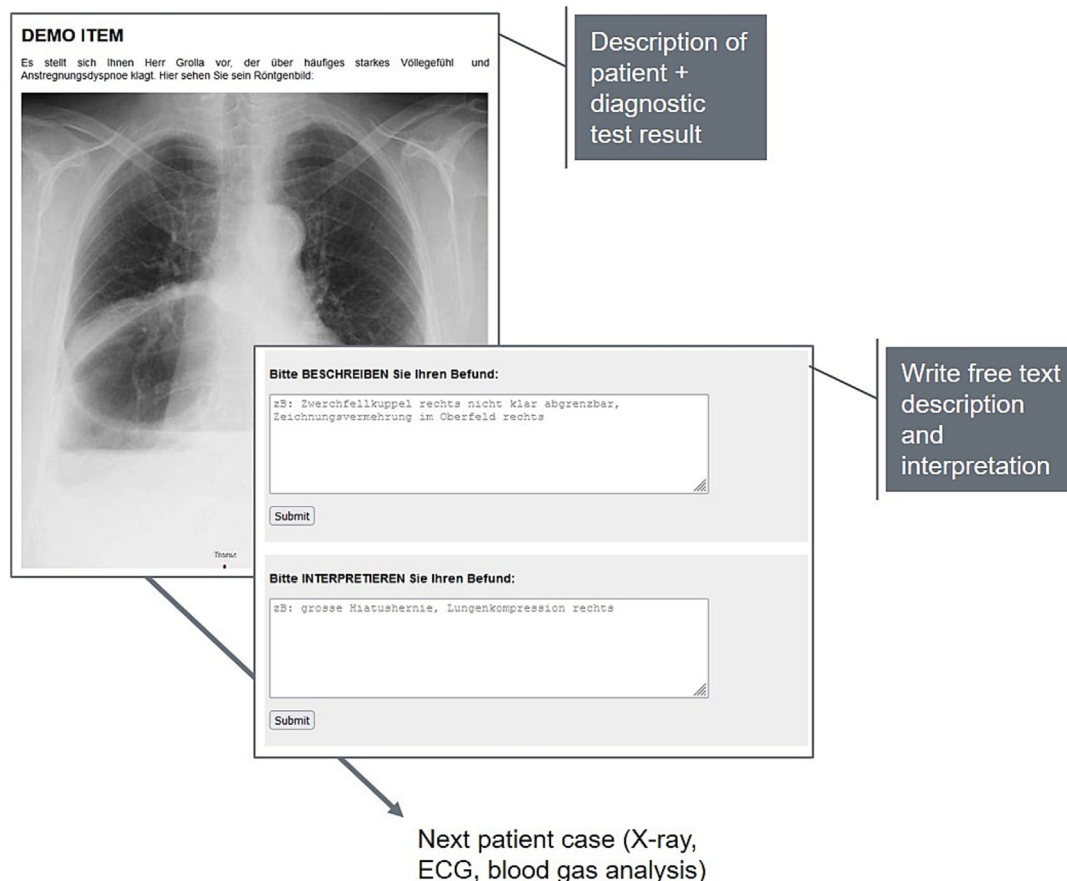
In the analysis task (Figure 2), the participants were shown  $N = 17$  diagnostic test results (i.e., six ECGs, six X-rays, and five laboratory results of blood samples) together with a one-sentence description of the patient (e.g., “A 44-year-old man with shortness of breath comes to you in the emergency room. Here, you can see his chest X-ray.”). For the exact wording of all items, please refer to our online supporting information (Kämmer et al., 2022). For each item, they were asked (a) to describe the pathological finding (e.g., location and aspect) and (b) to interpret it (e.g., its pathophysiological explanation) in a free text format.

In the integration task (Figure 3), the participants read  $N = 8$  vignettes of patients with respiratory distress, that is, short case descriptions that included previously interpreted diagnostic test results (e.g., “A 56-year-old woman with progressive shortness of breath over the last few days comes to you in the emergency room. The ECG is suggestive of right heart load and the chest X-ray shows a small infiltrate in the right lower lobe. ...”). They were then asked to indicate their most likely diagnosis in a free text format and to rate their related confidence (4-point Likert scale). Afterward, the participants were asked to select the diagnosis that most resembled their free text diagnosis from a provided list of six case-specific alternative diagnoses. This step was technically necessary to elicit their answer to the last request, in which the participants were presented with an alternative diagnosis from the same list of six diagnoses and asked to indicate the reasons for not selecting this alternative diagnosis in a free text format.

**TABLE 2** Participant demographics.

Variable	Condition		t test
	Individual ( $N = 39$ )	Pair ( $N = 70$ )	
Age (years), mean ( $SD$ )	22.77 (2.25)	23.46 (3.1)	$t(107) = -1.218, p = .226$
Gender ( $n$ )			
Male	19 (49%)	23 (33%)	
Female	20 (51%)	47 (67%)	
Semester, mean ( $SD$ )	7.46 (0.55)	7.76 (0.5)	$t(107) = 0.042, p = .966$
Number of emergency courses, mean ( $SD$ )	1.62 (0.49)	1.79 (0.51)	$t(107) = -1.696, p = .092$
Accuracy in knowledge test, mean % ( $SD$ )	12.46 (2.46)	12.57 (2.37)	$t(107) = -0.229, p = .819$





**FIGURE 2** Experimental procedure in the information-analysis task.

### 3.1 | Materials

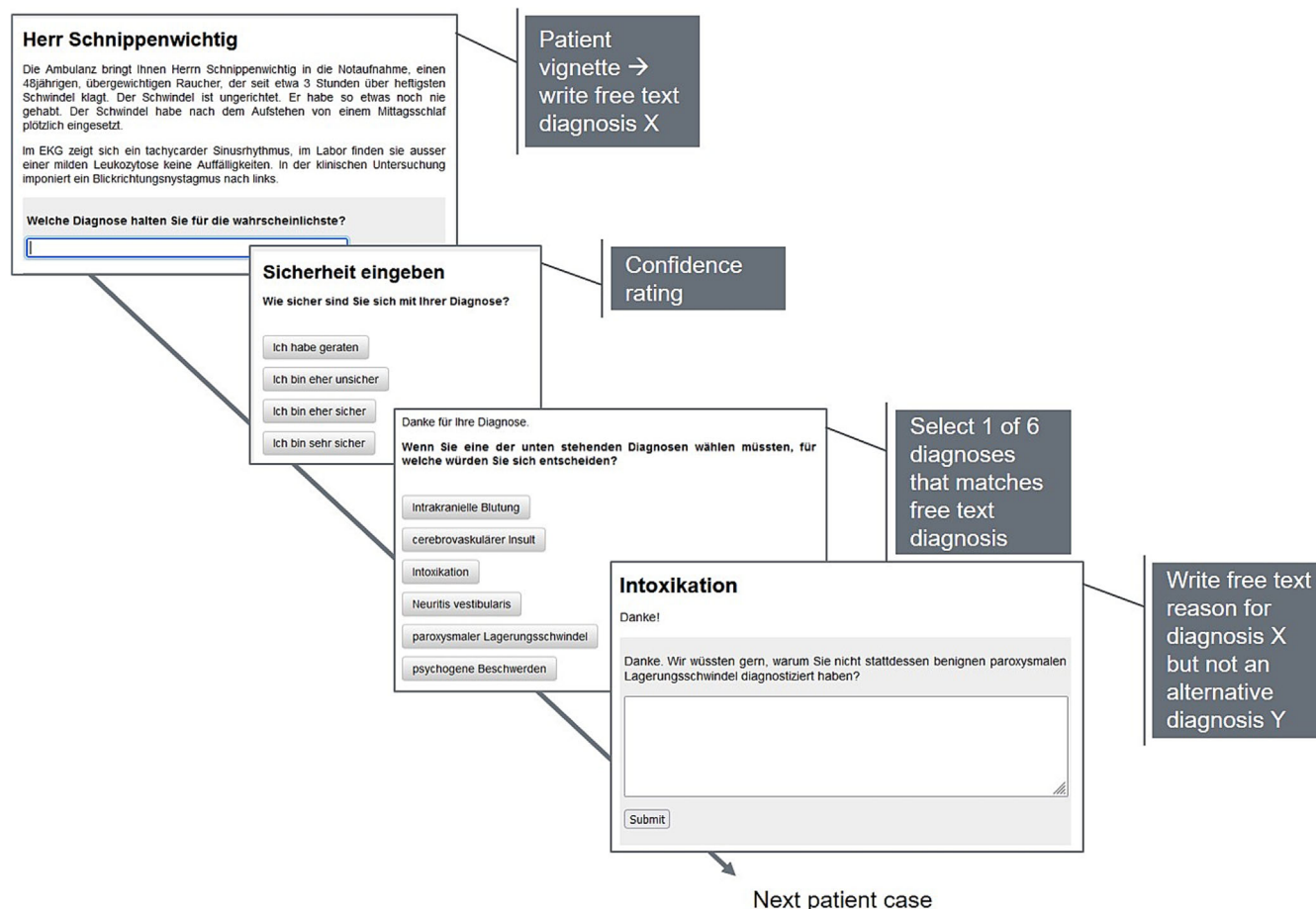
All items were limited to the topic of respiratory distress and chest pain and were made up using different books and collections as sources of inspiration and visual material (Trappe & Schuster, 2020; Woermann, 2000) as well as the clinical experience and examples of patients with typical findings in the emergency department in which the last author works. Items were pretested with a random sample of participants drawn from the same population as in the main study to check feasibility and intelligibility. The materials were then prepared using OpenLabyrinth (<http://openlabyrinth.ca>) to be presented digitally to the participants of the study.

### 3.2 | Measures

For the analysis task, accuracy ratings (1 = *correct*, 0.5 = *partially correct*, or 0 = *incorrect*) were obtained for the free text descriptions and interpretations from two trained physicians, blinded to the condition (see Table 3 for an overview of measures). We analyzed the accuracy ratings on three of the 17 items to assess the agreement between raters. We calculated intraclass correlation coefficients (ICCs), using the two-way random effect models and “single rater” unit, ICC(2,1),

agreement. There was good to excellent absolute agreement between the two raters (Cicchetti, 1994), with ICC = .73 (95% confidence interval, CI [0.66, 0.79]) for the descriptions and ICC = .84 (95% CI [0.79, 0.87]) for the interpretations. To obtain an overall measure of the accuracy of information analysis, we calculated the average accuracy ratings across the description and interpretation for each participant and item. This was done because these two steps are essentially two sides of the same coin and need to be considered together to provide a comprehensive assessment of the accuracy of information analysis. Difficulty of items ranged from very difficult ( $M_{\text{accuracy}} = 0.16$ ) to very easy ( $M_{\text{accuracy}} = 0.9$ ). An overview of all items and item difficulty can be found in the online supporting information (Kämmer et al., 2022).

Further, the free text descriptions and interpretations were rated concerning their complexity. Specifically, the complexity of descriptions was rated as either simple (0) or complex (1), on the basis of the wording, choice of terms, and accuracy of the language (Schmidt et al., 1990). Similarly, interpretations were marked as being either simple (0) or complex (1), without taking into account their correctness. As complexity ratings had to be based on the same free text as accuracy ratings, we took several precautions to minimize potential mutual influence including providing an example (see Table 3) and clear descriptions and definitions of the rating



**FIGURE 3** Experimental procedure in the information-integration task.

categories. Interrater agreement for three of 17 items was good, with  $ICC = .74$  (95% CI [0.65, 0.80]) for the descriptions and  $ICC = .63$  (95% CI [0.55, 0.71]) for the interpretations. As we did for the accuracy measure, we combined the complexity measures of descriptions and interpretations by averaging them per participant and item.

For the integration task, the accuracy of all free text diagnoses ( $N = 335$  unique diagnoses across the eight items) was rated by two trained physicians, blinded to the condition ( $1 = \text{correct}$ ,  $0.5 = \text{partially correct}$ , or  $0 = \text{incorrect}$ ). Interrater reliability was excellent with  $ICC = .93$  (95% CI [0.92, 0.94]). The arguments given as a reason for or against selecting a certain diagnosis were counted and the complexity of the reasons rated as simple (0), intermediate (0.5), or complex (1). We analyzed the complexity ratings on 1 of the eight cases to assess the agreement between raters. Interrater agreement was excellent with  $ICC = .85$  (95% CI [0.77, 0.90]).

### 3.3 | Analyses

To analyze the impact of collaboration on the (a) accuracy and (b) complexity of reasoning in the information-analysis

and information-integration subtasks, we fitted four successively more complex models using linear mixed-effects models with the dependent variables (a) accuracy and (b) complexity. Per dependent variable, we started with a baseline model that included the participant ID (per individual and pair), item ID as random intercepts, and gender (male vs. female) and modality of data collection (face-to-face vs. virtual) as fixed effects. In addition, we controlled for complexity when accuracy was the dependent variable, and for accuracy when complexity was the dependent variable, by including complexity and accuracy as a fixed effect, respectively. In Model 1, we additionally entered the condition (individual vs. pair) as a fixed effect. In Model 2, we additionally entered the task (analysis vs. integration) as a fixed effect. Finally, in Model 3, we additionally entered the interaction term Condition (individual vs. pair) \* Task (analysis vs. integration) as a fixed effect. We used the procedures provided in the lme4 package for fitting the linear mixed-effects models (Bates et al., 2014) in R (R Core Team, 2018). A required sample size of  $n = 34$  per condition was determined using G\*Power 3.1.9.7 (Faul et al., 2007) for repeated measures analyses of variance, assuming a small effect size,  $\alpha = .05$ , and  $\beta = .95$ .

To explore the advantages of collaboration over individual decision-making, studies have traditionally compared real pairs not



**TABLE 3** Overview of measures.

Dependent variable	Values	Explanation and examples
<b>Analysis task</b>		
Accuracy of description	0 = incorrect 0.5 = partially correct 1 = correct	Partially correct (0.5) when only one of many pathologies found or when correct and also incorrect findings were given
Accuracy of interpretation	0 = incorrect 0.5 = partially correct 1 = correct	
Complexity of description	0 = simple 1 = complex	How complex the description was, for example, focused on only a single pathological finding (0) versus systematic assessment (1)
Complexity of interpretation	0 = simple 1 = complex	How complex or specific the interpretation was, for example, pneumonia (0) versus pneumonia of the right upper lobe (1)
<b>Integration task</b>		
Accuracy of free text diagnosis	0 = incorrect 0.5 = partially correct 1 = correct	Partially correct when only an aspect of the diagnosis was given (e.g., left heart failure); correct when the full diagnosis was given (e.g., hypertensive pulmonary edema due to left heart failure)
Number of arguments	Range 1– <i>n</i>	An argument was defined as a single fact, such as a laboratory value, regardless of whether the argument was for or against the diagnosis. For example, “D-dimers and heart enzymes spoke for the diagnosis” counted as two arguments
Complexity of reason	0 = simple 0.5 = intermediate 1 = complex	How complex the reasoning for a selected diagnosis was, for example, list of reasons (0) versus simple reasoning (0.5) versus complex reasoning, for example, including pathomechanism (1)

Note: In addition, confidence ratings were obtained and the Cognitive Reflection Test (Frederick, 2005) was administered but neither was analyzed for the purpose of this study.

only to the average individual but also to the best member of nominal groups (Hautz et al., 2015; Larson, 2010; Laughlin et al., 2006; Wolf et al., 2015). In our study, we determined the best member in two ways: namely, (a) a priori, on the basis of the knowledge test administered at the start of the experiment, and (b) on an item-by-item basis. The latter process mimics the “truth wins” principle where one member recognizes the correct solution and demonstrates its correctness to the other member, who adopts the decision (Laughlin et al., 1991). By doing this, we aimed to assess whether real pairs exhibit strong synergy by outperforming the best individuals (Larson, 2010).

To compare real pairs with nominal pairs, we created all 741 unique nominal pairs from the 39 participants in the individual condition. We then identified the responses of the best member per nominal pair using methods (a) and (b). Specifically, for (a), we identified the member with the higher knowledge-test score per pair and assessed the accuracy of this individual's responses for all items. For (b), we determined the maximum accuracy per item per pair. We report the mean and distribution of accuracy of these simulations. All data are available on our OSF repository (Kämmer et al., 2022).

### 3.4 | Ethics

The ethics committee of the Canton of Bern deemed the study to be exempt from full ethical review (Req-2016-00330) because it did not

involve patients. Participation was voluntary and informed consent was obtained.

## 4 | RESULTS

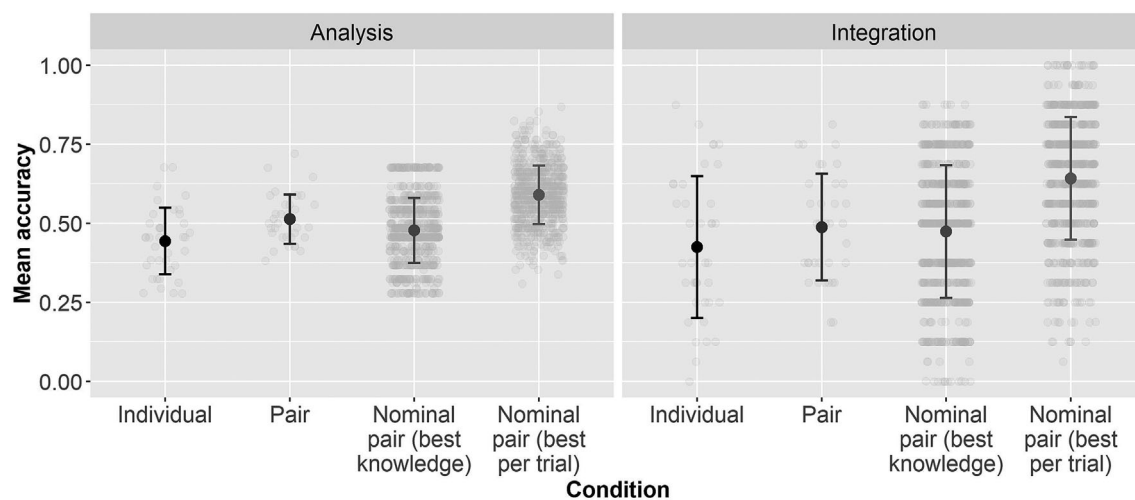
### 4.1 | Effects of teamwork on accuracy

In the information-analysis task, individuals achieved an accuracy of  $M = 0.44$  ( $SD = .11$ ) and pairs an accuracy of  $M = 0.51$  ( $SD = 0.08$ ; see Figure 4). In the information-integration task, individuals achieved an accuracy of  $M = 0.43$  ( $SD = 0.23$ ) and pairs an accuracy of  $M = 0.49$  ( $SD = 0.17$ ; see Figure 4). For pairs, higher accuracy was accompanied by a smaller set of unique incorrect diagnoses compared to results for individuals; in other words, collaboration decreased the diversity of incorrect diagnoses (similar to findings in Navajas et al., 2018; see online supporting information).

Results for the mixed-effects models can be seen in Table 4; in the following, we highlight the main findings.

In our baseline model, no main effects of gender ( $B = -0.02$ , 95% CI  $[-0.06, 0.02]$ ,  $p = .416$ ) or modality ( $B = 0.03$ , 95% CI  $[-0.01, 0.06]$ ,  $p = .190$ ) were revealed. However, an effect of complexity was revealed ( $B = 0.28$ , 95% CI  $[0.23, 0.33]$ ,  $p < .001$ ), indicating that complexity was higher with more accurate responses.

When we next included condition as a fixed effect (Model 1), results revealed a statistically significant increase in accuracy in the pair condition across tasks of an average of 5 percentage points



**FIGURE 4** Mean accuracy per task and condition (error bars  $\pm 1$  SD), with raw data in the background.

**TABLE 4** Results from four different linear mixed-effects models with the dependent variable accuracy.

Effect	Baseline model			Model 1			Model 2			Model 3		
	B	95% CI	p value	B	95% CI	p value	B	95% CI	p value	B	95% CI	p value
Fixed effects												
(intercept)	0.31	0.20–0.42	<b>&lt;.001</b>	0.28	0.17–0.39	<b>&lt;.001</b>	0.25	0.12–0.38	<b>&lt;.001</b>	0.25	0.12–0.39	<b>&lt;.001</b>
Complexity <sup>a</sup>	0.28	0.23–0.33	<b>&lt;.001</b>	0.28	0.23–0.33	<b>&lt;.001</b>	0.28	0.23–0.33	<b>&lt;.001</b>	0.28	0.23–0.33	<b>&lt;.001</b>
Modality <sup>b</sup>	0.03	–0.01–0.06	.190	0.03	–0.01–0.06	.175	0.03	–0.01–0.06	.175	0.03	–0.01–0.06	.175
Gender <sup>c</sup>	–0.02	–0.06–0.02	.416	–0.01	–0.04–0.03	.731	–0.01	–0.04–0.03	.731	–0.01	–0.04–0.03	.730
Condition <sup>d</sup>				0.05	0.01–0.09	<b>.011</b>	0.05	0.01–0.09	<b>.011</b>	0.04	0.00–0.08	<b>.046</b>
Task <sup>e</sup>							0.09	–0.13–0.31	.443	0.08	–0.14–0.30	.491
Condition * task										0.02	–0.05–0.09	.573
Random effects												
$\sigma^2$	0.10			0.10			0.10			0.10		
$\tau_{00}$												
Participant ID	0.00			0.00			0.00			0.00		
Item ID	0.07			0.07			0.07			0.07		
Marginal $R^2$ / conditional $R^2$	.078/.454			.082/.454			.064/.447			.064/.447		

Note:  $N = 1704$  observations in each model.  $N = 73$  participant IDs (one ID per individual and pair), and  $N = 25$  item IDs in each model. CI = confidence interval.  $p$ -values  $p < .050$  written in bold.

<sup>a</sup>0 (simple) to 1 (complex).

<sup>b</sup>Face-to-face versus online.

<sup>c</sup>Male versus female.

<sup>d</sup>Individual versus pair.

<sup>e</sup>Analysis versus integration.

( $B = 0.05$ , 95% CI [0.01, 0.09],  $p = .011$ ). When we additionally entered task as a fixed effect in Model 2, no additional effect of type of task was revealed. Last, in Model 3, we included the interaction term Condition \* Task as a fixed effect. The main effect of condition remained significant ( $B = 0.04$ , 95% CI [0.00, 0.08],  $p = .046$ ), but there was neither a main effect of the type of task ( $B = 0.08$ , 95% CI [–0.14, 0.30],  $p = .491$ ) nor an interaction effect with the type of task ( $B = 0.02$ , 95% CI [–0.05, 0.09],  $p = .573$ ; see Table 4).

## 4.2 | Benchmarking the accuracy of real pairs against nominal pairs

Next, we tested whether real pairs performed at or above the level of the best member of pairs constructed from the participants in the individual condition, using computer simulations. Analyses revealed that (a) the more knowledgeable member in nominal pairs achieved an accuracy of  $M = 0.48$  (95% CI [0.47, 0.49]) in the information-analysis

task and an accuracy of  $M = 0.47$  (95% CI [0.46, 0.49]) in the information-integration task (see Figure 4). Further, (b) the best member (on an item-by-item basis) achieved an accuracy of  $M = 0.59$  (95% CI [0.58, 0.60]) in the information-analysis task and an accuracy of  $M = 0.64$  (95% CI [0.63, 0.66]) in the information-integration task. These results indicate that real pairs performed at the level of their more knowledgeable member but below the best possible member (on an item-by-item basis) of nominal pairs.

### 4.3 | Effect of teamwork on the complexity of reasoning

In the information-analysis task, the average combined complexity of descriptions and specificity of interpretations provided by individuals was  $M = 0.63$  ( $SD = 0.20$ ) and by pairs,  $M = 0.71$  ( $SD = 0.11$ ; see Figure 5). In the information-integration task, the complexity of reasons given by individuals was on average  $M = 0.17$  ( $SD = 0.17$ ) and by pairs,  $M = 0.13$  ( $SD = 0.19$ ). On average, individuals provided  $M = 1.78$  ( $SD = 1.05$ ) and pairs  $M = 2.1$  ( $SD = 1.18$ ) reasons against the unchosen diagnosis.

We again fitted successively more complex models using linear mixed-effects models with the dependent variable complexity. Results can be seen in Table 5; in the following, we highlight the main findings.

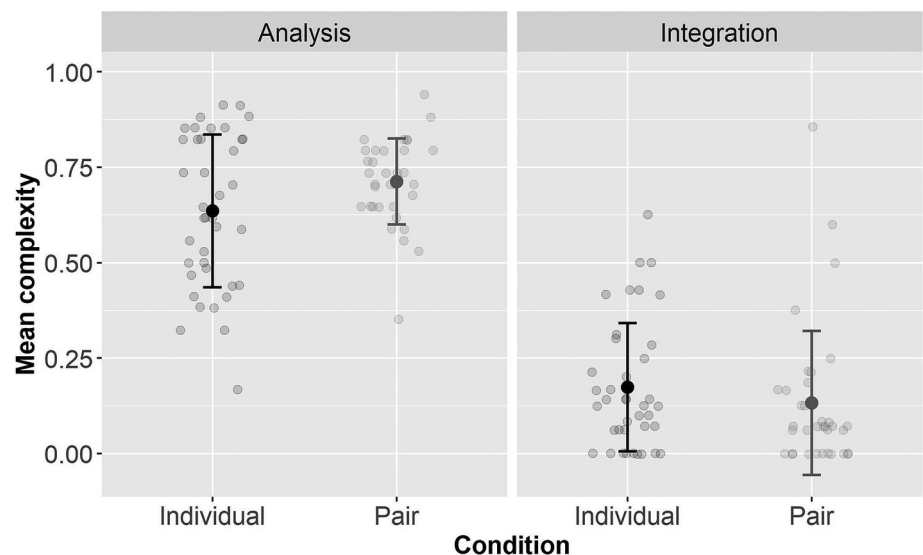
In our baseline model, no main effects of gender ( $B = -0.02$ , 95% CI [-0.08, 0.04],  $p = .585$ ) or modality ( $B = 0.03$ , 95% CI [-0.03, 0.08],  $p = .387$ ) were revealed. However, an effect of accuracy was revealed ( $B = 0.22$ , 95% CI [0.17, 0.26],  $p < .001$ ), indicating that accuracy was higher with more complex responses. When we included condition as a fixed effect in Model 1, no effect of condition was revealed ( $B = 0.05$ , 95% CI [0.01, 0.09],  $p = .011$ ). When we additionally entered task as a fixed effect in Model 2, a main effect of type of task was revealed, indicating that complexity in the information-analysis task was higher than in the information-integration task ( $B = -0.51$ , 95% CI [-0.62, -0.39],

$p < .001$ ). This effect remained significant when we additionally included the interaction term Condition \* Task as a fixed effect in Model 3 ( $B = -0.45$ , 95% CI [-0.57, -0.34],  $p < .001$ ). In addition, the interaction effect was significant ( $B = -0.12$ , 95% CI [-0.18, -0.06],  $p < .001$ ). The interaction plot indicates that pairs provided more complex descriptions and interpretations than individuals in the information-analysis task but less complex reasons for their diagnoses in the information-integration task; yet, post-hoc simple main effect tests showed no significant main effect of condition (both  $p \geq .177$ ).

## 5 | DISCUSSION

When do teams outperform individuals? This question has a long tradition in psychological research (e.g., Watson, 1928; for a review see Mathieu et al., 2017) and continues to be relevant, as many modern organizations rely on teams today. We contribute to the research on this important question by exploring the performance benefits of teams over individuals in the context of the diagnostic decision-making process. For this, we studied the impact of collaboration on performance separately in the information-analysis and information-integration subtasks that are part of the diagnostic decision-making process. We employed a high-fidelity task scenario of practical relevance with participants having to apply their medical expertise to the task. Our focus on the subtasks instead of the complete diagnostic task was driven by theoretical and practical considerations: In line with Larson (2010), we think that research on the subtask structure is necessary to understand overall performance (differences) better. And in applied settings, such as the hospital setting, opportunities for direct collaboration are often limited so that recommendations for when collaboration yields the highest benefits would be particularly valuable.

Our analyses revealed that collaborating pairs outperformed individuals in the information-analysis and information-integration



**FIGURE 5** Mean complexity per task and condition (error bars  $\pm 1$  SD), with raw data in the background.

**TABLE 5** Results from four different linear mixed-effects models with the dependent variable complexity.

Effect	Baseline model			Model 1			Model 2			Model 3		
	B	95% CI	p value	B	95% CI	p value	B	95% CI	p value	B	95% CI	p value
Fixed effects												
(intercept)	0.40	0.28–0.52	<b>&lt;.001</b>	0.39	0.26–0.51	<b>&lt;.001</b>	0.55	0.46–0.64	<b>&lt;.001</b>	0.53	0.44–0.62	<b>&lt;.001</b>
Accuracy <sup>a</sup>	0.22	0.17–0.26	<b>&lt;.001</b>	0.22	0.17–0.26	<b>&lt;.001</b>	0.22	0.18–0.26	<b>&lt;.001</b>	0.22	0.18–0.26	<b>&lt;.001</b>
Modality <sup>b</sup>	0.03	–0.03–0.08	.387	0.03	–0.03–0.08	.385	0.03	–0.03–0.08	.391	0.02	–0.03–0.08	.411
Gender <sup>c</sup>	–0.02	–0.08–0.04	.585	–0.01	–0.07–0.05	.700	–0.01	–0.07–0.05	.701	–0.01	–0.07–0.05	.708
Condition <sup>d</sup>				0.02	–0.04–0.08	.447	0.02	–0.04–0.08	.448	0.06	–0.00–0.12	.061
Task <sup>e</sup>							–0.51	–0.62–0.39	<b>&lt;.001</b>	–0.45	–0.57 to –0.34	<b>&lt;.001</b>
Condition * task										–0.12	–0.18 to –0.06	<b>&lt;.001</b>
Random effects												
$\sigma^2$	0.08			0.08			0.08			0.08		
$\tau_{00}$												
Participant ID	0.01			0.01			0.01			0.01		
Item ID	0.07			0.07			0.02			0.02		
Marginal $R^2$ / conditional $R^2$	.050/.538			.052/.539			.370/.536			.374/.541		

Note:  $N = 1704$  observations in each model.  $N = 73$  participant IDs (one ID per individual and pair), and  $N = 25$  item IDs in each model. CI = confidence interval.  $p$ -values  $\leq .050$  written in bold.

<sup>a</sup>0–1.

<sup>b</sup>Face-to-face versus online.

<sup>c</sup>Male versus female.

<sup>d</sup>Individual versus pair.

<sup>e</sup>Analysis versus integration.

subtasks. Further, we found that the performance benefits of pairs over individuals were of a similar margin in both subtasks, yet, only of small size (approximately six percentage points per task,  $\beta = .14$ ). Whether this similarity in effect size between subtasks is a sign of the same underlying mechanisms, such as the role of memory or knowledge base, needs to be addressed with further research. Our task analysis (see Section 1 and Table 1) may help researchers develop hypotheses for future investigations.

When comparing real pairs with nominal pairs (i.e., the same number of independent individuals), we found that real pairs performed at the level of the most knowledgeable member but below that of the most accurate member (on an item-by-item basis), suggesting that collaboration was only as effective as selecting the better member a priori (determined by a knowledge pretest) of a group, and not beyond (i.e., thus demonstrating only a weak synergistic performance gain; cf. Larson, 2010).

We also found correct answers to be more complex than incorrect answers, indicating that more knowledge was applied or more of the available information was integrated when an answer was correct. Yet, complexity did not differ per se between individuals and pairs, indicating that neither more nor more diverse knowledge is being combined during collaboration (Vollrath et al., 1989). Complexity thus seems to be an indicator of accuracy rather than of collaboration.

One conclusion from these results could be that the performance benefits of pairs, which we had previously observed when pairs worked together on the complete diagnostic process

(Hautz et al., 2015), have their origin in both subtasks (the information-gathering subtask was ruled out in the previous study). Whether the benefits of collaboration accumulate if several subtasks are attended to together remains an open question for future studies that use a different experimental design.

From a practical point of view, selective collaboration during single phases of the diagnostic process arguably has a number of benefits, including accuracy gains (as we show here), and yet only limited additional coordination requirements, time, and personnel resources. Collaboration during the complete process necessarily requires more resources; yet, we expect that it also yields higher performance gains, possibly for the following reasons:

First, breaking down a task and engaging in only selective collaboration during individual subtasks could potentially lead to a reduction in task complexity. However, it is plausible that only highly complex tasks truly unleash the full potential of teams, thus leading to strong synergy effects. Second when teams collectively attend to a complete case, the subtasks that were individually explored in the current study are likely to be performed multiple times, presenting an opportunity for additive effects to occur.

Third, as highlighted by Larson (2010, p. 41), “when there are multiple subtasks to perform, a decision must be made about the order in which they will be done.” This becomes particularly crucial in the diagnostic process, where one test result can influence the selection of the next test, and integrating available information may reveal the necessity to revisit previous results that do not align with the

overall picture. In essence, not only are multiple subtasks involved, such as information gathering, analysis, and integration, but there is also an orthogonal subtask of sequencing these subtasks. The quality of this sequencing process likely impacts the overall task performance. By engaging in selective collaboration, teams might miss out on leveraging their strengths during these critical sequencing decisions.

Fourth, teams have the advantage of resource sharing and the ability to assign different subtasks to members based on their capabilities. For instance, a team may identify members with expertise in interpreting X-rays or analyzing ECGs. As a result, the group can choose to weight each member's input according to their (stated) capability (i.e., differential weighting; Sniezek & Henry, 1989, 1990). Larson (2010, p. 7) suggested that in practical real-world settings, teams can optimize performance by delegating subtasks to members who possess the greatest skills for performing those subtasks, thus achieving an optimal person-subtask fit and harnessing synergy by appropriately weighting opinions.

## 5.1 | Limitations and open questions

Our study comes with some limitations with respect to the generalizability of results concerning group size, heterogeneity, and expertise level. First, although studying pairs allows one to study “pure” effects of collaboration on information processing with members having to spend minimal effort on additional coordination demands as is the case in larger teams, it leaves the question open of how effects scale up with group size and whether processing benefits may still outweigh coordination demands in larger teams.

Second, although studying advanced medical students is of practical relevance, as they still have important knowledge gaps (Braun et al., 2017) and may thus benefit most from any support, it remains an open question how more experienced diagnosticians might benefit from targeted collaboration. Third, we studied rather homogeneous pairs with regard to their experience level, gender, profession, specialty, and hierarchy; it is thus a task for future research to study the impact of collaboration in more heterogeneous groups, such as when the knowledge overlap is smaller and there is potentially more to gain through knowledge exchange. Also, future studies could extend our study to other common forms of interaction such as cross-checking (Freund et al., 2018) or advice-based decision-making (Kämmer et al., 2023; Schultze & Loschelder, 2021).

Fourth, we studied the effect of synchronous face-to-face interaction; yet, in real-world health-care settings, professionals often interact without verbal or face-to-face interaction and more through reading others' notes (Olson et al., 2020). Future research could thus compare the effect of different forms of interaction on the diagnostic process and its subtasks.

Fifth, we studied the immediate effect of collaboration on the task product (i.e., accuracy of the analysis or diagnosis); yet, particularly in an educational setting such as a teaching hospital, more long-term learning effects are another relevant outcome variable that could be studied in future research (Shanks et al., 2013).

Last, we studied short-term collaboration in ad hoc teams without time restrictions; repeated interactions of a longer period, which may be common in some health-care settings, may lead to emergent shared knowledge structures (Kozlowski & Ilgen, 2006) and thus to greater benefits through collaboration.

## 5.2 | Conclusion

How can decision-making under uncertainty be improved, particularly when wrong decisions may have detrimental consequences, as is often the case in medicine? Our findings suggest that collaboration during the subtasks information analysis and information integration during the diagnostic decision-making process yield small benefits when compared to working alone. Further research is needed to uncover the mechanisms underlying the benefits of collaboration during (parts of) the decision process.

## ACKNOWLEDGMENTS

The authors are grateful to Fabian Stroben, Anna Wittenstein, and Tobias Bolte for their support with data collection. We thank Anita Todd for language editing as well as Stefan Schulz-Hardt for valuable suggestions concerning the data analysis. We also thank Margarete Boos for her comments on an earlier version of the manuscript and Jim Larson as well as an anonymous reviewer for their constructive reviews.

JEK has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 894536, project “TeamUp” and under the grant agreement no. 101021775, project “Med1stMR.” Preliminary results were presented at the 64th Conference of Experimental Psychologists (TeaP), the 17th Annual INGroup Conference, and the 52nd Conference of the German Society for Psychology in 2022. Open access funding provided by Universität Bern.

## DATA AVAILABILITY STATEMENT

Data are available under <https://doi.org/10.17605/OSF.IO/EA5BP>.

## ORCID

Juliane E. Kämmer  <https://orcid.org/0000-0001-6042-8453>

Stefan K. Schaubert  <https://orcid.org/0000-0002-1832-2732>

Stefanie C. Hautz  <https://orcid.org/0000-0003-4715-8465>

Wolf E. Hautz  <https://orcid.org/0000-0002-2445-984X>

## REFERENCES

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv*, 1–51. <https://doi.org/10.48550/arXiv.1406.5823>
- Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, 121(5), S2–S23. <https://doi.org/10.1016/j.amjmed.2008.01.001>
- Boos, M., Schauenburg, B., Strack, M., & Belz, M. (2013). Social validation of shared and nonvalidation of unshared information in group



- discussions. *Small Group Research*, 44(3), 257–271. <https://doi.org/10.1177/1046496413484068>
- Braun, L. T., Zwaan, L., Kiesewetter, J., Fischer, M. R., & Schmidmaier, R. (2017). Diagnostic errors by medical students: Results of a prospective qualitative study. *BMC Medical Education*, 17(1), 191. <https://doi.org/10.1186/s12909-017-1044-7>
- Chalos, P., & Pickard, S. (1985). Information choice and cue use: An experiment in group information processing. *Journal of Applied Psychology*, 70(4), 634–641. <https://doi.org/10.1037/0021-9010.70.4.634>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290.
- Collins, B. E., & Guetzkow, H. S. (1964). *A social psychology of group processes for decision-making*. Wiley.
- Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, & The National Academies of Sciences, Engineering, and Medicine. (2015). In E. P. Balogh, B. T. Miller, & J. R. Ball (Eds.), *Improving diagnosis in health care*. National Academies Press. <https://doi.org/10.17226/21794>
- De Dreu, C. K. W., Nijstad, B. A., & van Knippenberg, D. (2008). Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review*, 12(1), 22–49. <https://doi.org/10.1177/1088868307304092>
- DeChurch, L. A., Burke, C. S., Shuffler, M. L., Lyons, R., Doty, D., & Salas, E. (2011). A historiometric analysis of leadership in mission critical multi-team environments. *The Leadership Quarterly*, 22(1), 152–169. <https://doi.org/10.1016/j.leaqua.2010.12.013>
- Deloitte Insights. (2019). 2019 Deloitte global human capital trends report. Deloitte Touche Tohmatsu Limited. [https://www2.deloitte.com/content/dam/insights/us/articles/5136\\_HC-Trends-2019/DI\\_HC-Trends-2019.pdf](https://www2.deloitte.com/content/dam/insights/us/articles/5136_HC-Trends-2019/DI_HC-Trends-2019.pdf)
- Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3), 497–509. [https://doi.org/10.1016/S0022-1031\(03\)00040-4](https://doi.org/10.1016/S0022-1031(03)00040-4)
- Edmondson, A. C. (2012). *Teaming: How organizations learn, innovate, and compete in the knowledge economy*. Jossey-Bass.
- Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes*, 73(2–3), 116–141. <https://doi.org/10.1006/obhd.1998.2758>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. <https://doi.org/10.3758/BF03193146>
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Freund, Y., Goulet, H., Leblanc, J., Bokobza, J., Ray, P., Maignan, M., Guinemer, S., Truchot, J., Féral-Pierssens, A.-L., Yordanov, Y., Philippon, A.-L., Rouff, E., Bloom, B., Cachanado, M., Rousseau, A., Simon, T., & Riou, B. (2018). Effect of systematic physician cross-checking on reducing adverse events in the emergency department: The CHARMED cluster randomized trial. *JAMA Internal Medicine*, 178(6), 812–819. <https://doi.org/10.1001/jamainternmed.2018.0607>
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62(1), 451–482. <https://doi.org/10.1146/annurev-psych-120709-145346>
- Graber, M. L., Rusz, D., Jones, M. L., Farm-Franks, D., Jones, B., Gluck, J. C., Thomas, D. B., Gleason, K. T., Welte, K., & Abfalter, J. (2017). The new diagnostic team. *Diagnosis*, 4(4), 225–238. <https://doi.org/10.1515/dx-2017-0022>
- Hagemann, V., Kluge, A., & Ritzmann, S. (2011). High responsibility teams—Eine systematische Analyse von Teamarbeitskontexten für einen effektiven Kompetenzerwerb. *Journal Psychologie Des Alltags*, 4(1), 22–42.
- Hautz, W. E., Kämmer, J. E., Hautz, S. C., Sauter, T. C., Zwaan, L., Exadaktylos, A. K., Birrenbach, T., Maier, V., Müller, M., & Schaubert, S. K. (2019). Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 27(1), 54. <https://doi.org/10.1186/s13049-019-0629-z>
- Hautz, W. E., Kämmer, J. E., Schaubert, S. K., Spies, C. D., & Gaissmaier, W. (2015). Diagnostic performance by medical students working individually or in teams. *Jama*, 313(3), 303–304. <https://doi.org/10.1001/jama.2014.15770>
- Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology*, 59(4), 705–718. <https://doi.org/10.1037/0022-3514.59.4.705>
- Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, 121(1), 43–64. <https://doi.org/10.1037/0033-2909.121.1.43>
- Hirokawa, R. Y., & Pace, R. (1983). A descriptive investigation of the possible communication-based reasons for effective and ineffective group decision making. *Communication Monographs*, 50(4), 363–379. <https://doi.org/10.1080/03637758309390175>
- Kämmer, J. E., Choshen-Hillel, S., Müller-Trede, J., Black, S. L., & Weibler, J. (2023). A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision*, 10, 107–137. <https://doi.org/10.1037/dec0000199>
- Kämmer, J. E., Ernst, K., Grab, K., Schaubert, S. K., Hautz, S. C., Penders, D., & Hautz, W. E. (2022). Material for “collaboration during the diagnostic decision-making process: When does it help?” OSF. <https://doi.org/10.17605/OSF.IO/EA5BP>
- Kämmer, J. E., Hautz, W. E., Herzog, S. M., Kunina-Habenicht, O., & Kurvers, R. H. J. M. (2017). The potential of collective intelligence in emergency medicine: Pooling medical students' independent decisions improves diagnostic performance. *Medical Decision Making*, 37(6), 715–724. <https://doi.org/10.1177/0272989X17696998>
- Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4), 681–706. <https://doi.org/10.1037/0022-3514.65.4.681>
- Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, 55(1), 623–655. <https://doi.org/10.1146/annurev.psych.55.090902.142009>
- Kiesewetter, J., Fischer, F., & Fischer, M. R. (2017). Collaborative clinical reasoning—A systematic review of empirical studies. *Journal of Continuing Education in the Health Professions*, 37(2), 123–128. <https://doi.org/10.1097/CEH.0000000000000158>
- Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, 7(3), 77–124. <https://doi.org/10.1111/j.1529-1006.2006.00030.x>
- Krupat, E., Wormwood, J., Schwartzstein, R. M., & Richards, J. B. (2017). Avoiding premature closure and reaching diagnostic accuracy: Some key predictive factors. *Medical Education*, 51(11), 1127–1137. <https://doi.org/10.1111/medu.13382>
- Kunina-Habenicht, O., Hautz, W. E., Knigge, M., Spies, C., & Ahlers, O. (2015). Assessing clinical reasoning (ASCLIRE): Instrument development and validation. *Advances in Health Sciences Education*, 20(5), 1205–1224. <https://doi.org/10.1007/s10459-015-9596-y>
- Larson, J. R. (2010). *In search of synergy in small group performance*. Psychology Press.
- Larson, J. R., Christensen, C., Abbott, A. S., & Franz, T. M. (1996). Diagnosing groups: Charting the flow of information in medical decision-making teams. *Journal of Personality and Social Psychology*, 71(2), 315–330. <https://doi.org/10.1037/0022-3514.71.2.315>
- Larson, J. R., Christensen, C., Franz, T. M., & Abbott, A. S. (1998). Diagnosing groups: The pooling, management, and impact of shared and unshared case information in team-based medical decision making. *Journal of Personality and Social Psychology*, 75(1), 93–108. <https://doi.org/10.1037/0022-3514.75.1.93>



- Larson, J. R., Foster-Fishman, P. G., & Keys, C. B. (1994). Discussion of shared and unshared information in decision-making groups. *Journal of Personality and Social Psychology*, 67(3), 446–461. <https://doi.org/10.1037/0022-3514.67.3.446>
- 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(4), 644–651. <https://doi.org/10.1037/0022-3514.90.4.644>
- Laughlin, P. R., VanderStoep, S. W., & Hollingshead, A. B. (1991). Collective versus individual induction: Recognition of truth, rejection of error, and collective information processing. *Journal of Personality and Social Psychology*, 61(1), 50–67. <https://doi.org/10.1037/0022-3514.61.1.50>
- Lerner, J. S., & Tetlock, P. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2), 255–275. <https://doi.org/10.1037/0033-2909.125.2.255>
- Lorge, I., & Solomon, H. (1955). Two models of group behavior in the solution of eureka-type problems. *Psychometrika*, 20(2), 139–148. <https://doi.org/10.1007/BF02288986>
- Massaro, D. W., & Cowan, N. (1993). Information processing models: Microscopes of the mind. *Annual Review of Psychology*, 44(1), 383–425. <https://doi.org/10.1146/annurev.ps.44.020193.002123>
- Mathieu, J. E., Heffner, T., Goodwin, G., Salas, E., & Cannon-Bowers, J. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85(2), 273–283. <https://doi.org/10.1037/0021-9010.85.2.273>
- Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the *journal of applied psychology*. *Journal of Applied Psychology*, 102(3), 452–467. <https://doi.org/10.1037/apl0000128>
- Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 2(2), 126–132. <https://doi.org/10.1038/s41562-017-0273-4>
- Olson, A. P. J., Durning, S. J., Fernandez Branson, C., Sick, B., Lane, K. P., & Rencic, J. J. (2020). Teamwork in clinical reasoning—Cooperative or parallel play? *Diagnosis*, 7(3), 307–312. <https://doi.org/10.1515/dx-2020-0020>
- Petty, R. E., Harkins, S. G., & Williams, K. D. (1980). The effects of group diffusion of cognitive effort on attitudes: An information-processing view. *Journal of Personality and Social Psychology*, 38(1), 81–92. <https://doi.org/10.1037/0022-3514.38.1.81>
- Propp, K. M. (1999). Collective information processing in groups. In L. R. Frey, D. Gouran, & M. S. Poole (Eds.), *The handbook of group communication theory and research* (pp. 225–250). Sage Publications.
- R Core Team. (2018). *R: A language and environment for statistical computing [computer software]*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, 5(6), 649–663. <https://doi.org/10.1177/1745691610388763>
- Salas, E. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*, 50(3), 540–547. <https://doi.org/10.1518/001872008X288457>
- Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine*, 65(10), 611–621. <https://doi.org/10.1097/00001888-199010000-00001>
- Schultze, T., & Loschelder, D. D. (2021). How numeric advice precision affects advice taking. *Journal of Behavioral Decision Making*, 34(3), 303–310. <https://doi.org/10.1002/bdm.2211>
- Schulz-Hardt, S., Frey, D., Lüthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology*, 78(4), 655–669. <https://doi.org/10.1037/0022-3514.78.4.655>
- Schulz-Hardt, S., Jochims, M., & Frey, D. (2002). Productive conflict in group decision making: Genuine and contrived dissent as strategies to counteract biased information seeking. *Organizational Behavior and Human Decision Processes*, 88(2), 563–586. [https://doi.org/10.1016/S0749-5978\(02\)00001-8](https://doi.org/10.1016/S0749-5978(02)00001-8)
- Shanks, D., Brydges, R., den Brok, W., Nair, P., & Hatala, R. (2013). Are two heads better than one? Comparing dyad and self-regulated learning in simulation training. *Medical Education*, 47(12), 1215–1222. <https://doi.org/10.1111/medu.12284>
- Singh, H., Meyer, A. N. D., & Thomas, E. J. (2014). The frequency of diagnostic errors in outpatient care: Estimations from three large observational studies involving US adult populations. *BMJ Quality and Safety*, 23(9), 727–731. <https://doi.org/10.1136/bmjqs-2013-002627>
- Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, 43(1), 1–28. [https://doi.org/10.1016/0749-5978\(89\)90055-1](https://doi.org/10.1016/0749-5978(89)90055-1)
- Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes*, 45(1), 66–84. [https://doi.org/10.1016/0749-5978\(90\)90005-T](https://doi.org/10.1016/0749-5978(90)90005-T)
- Steiner, I. D. (1972). *Group process and productivity*. Academic Press.
- Trappe, H.-J., & Schuster, H.-P. (2020). *EKG-Kurs für Isabel: 380 - Abbildungen*. Georg Thieme Verlag. <https://doi.org/10.1055/b000000429>
- Turner, M. E., & Pratkanis, A. R. (1998). Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. *Organizational Behavior and Human Decision Processes*, 73(2–3), 105–115. <https://doi.org/10.1006/obhd.1998.2756>
- Vollrath, D. A., Sheppard, B. H., Hinsz, V. B., & Davis, J. H. (1989). Memory performance by decision-making groups and individuals. *Organizational Behavior and Human Decision Processes*, 43(3), 289–300. [https://doi.org/10.1016/0749-5978\(89\)90040-X](https://doi.org/10.1016/0749-5978(89)90040-X)
- Watson, G. B. (1928). Do groups think more efficiently than individuals? *The Journal of Abnormal and Social Psychology*, 23(3), 328–336. <https://doi.org/10.1037/h0072661>
- Weldon, M. S., Blair, C., & Huesch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6), 1568–1577. <https://doi.org/10.1037/0278-7393.26.6.1568>
- Woermann, U. (2000). *RadioSurf—Interaktive Lernmodule zur diagnostischen Radiologie [Computer software]*. University of Bern. <https://radiosurf.elearning.aum.unibe.ch/htmls/radskeleton.html>
- Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. J. M. (2015). Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PLoS ONE*, 10(8), e0134269. <https://doi.org/10.1371/journal.pone.0134269>
- Zwaan, L., de Bruijne, M., Wagner, C., Thijs, A., Smits, M., van der Wal, G., & Timmermans, D. R. (2010). Patient record review of the incidence, consequences, and causes of diagnostic adverse events. *Archives of Internal Medicine*, 170(12), 1015–1021. <https://doi.org/10.1001/archinternmed.2010.146>

**How to cite this article:** Kämmer, J. E., Ernst, K., Grab, K., Schaub, S. K., Hautz, S. C., Penders, D., & Hautz, W. E. (2024). Collaboration during the diagnostic decision-making process: When does it help? *Journal of Behavioral Decision Making*, 37(1), e2357. <https://doi.org/10.1002/bdm.2357>