Advances in Health Sciences Education (2020) 25:905–911
https://doi.org/10.1007/s10459-020-09956-z

COMMENTARY

Whether two heads are better than one is the wrong question (though sometimes they are)

Wolf E. Hautz[1] · **Stefanie C. Hautz**[1] · **Juliane E. Kämmer**[2,3]

Received: 11 August 2019 / Accepted: 13 January 2020 / Published online: 6 February 2020

A recent editorial by Norman (2019) in this journal asked whether "[t]wo heads are better than one". Following a light-hearted and insightful deliberation on medical training specifically and problem solving in general, either individually or in groups, Norman concluded that "two (independent) heads are better than one (group of two heads)" (2019). We applaud the author for questioning a widely accepted belief and for fostering a discussion on the pearls and pitfalls of collaboration in medicine. Drawing on a review of the medical and psychological literature, we would, however, argue that his conclusion (a) leaves important evidence out of consideration, (b) results from a conceptual oversimplification, and (c) addresses the wrong question. In the following, we highlight relevant research on the merits of one versus more heads in the context of medical diagnoses, present a theoretical conception of the problem, and conclude that the question of whether or not to collaborate should be replaced by that of when and why to collaborate, aggregate or work in isolation. We conclude with specific suggestions for further research in this area, illustrating our point with an example from research into group leadership.

What is the evidence?

As Norman (2019) notes, ample research on collective intelligence or the wisdom of the crowd shows that two (or more) independent heads are better than one: Algorithmically aggregating two or more opinions usually outperforms the average (e.g., Kämmer et al. 2017; Kurvers et al. 2015; Surowiecki 2005) and sometimes even the best *individual* (Wolf et al. 2015). The paper by Barnett et al. (2019), which sparked Norman's editorial and which he skillfully dissects, reports similar findings, albeit challenged by methodological limitations and an unusual use of the term "group" (Norman 2019).

✉ Wolf E. Hautz
wolf.hautz@insel.ch

[1] Department of Emergency Medicine, Inselspital University Hospital, University of Bern, Freiburgstrasse 16c, 3010 Bern, Switzerland

[2] Institute of Health and Nursing Science, Charité – Universitätsmedizin Berlin Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Oudenarder Str. 16, 13347 Berlin, Germany

[3] Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Given the relevance and ubiquity of teamwork today (Deloitte Insights 2019), the more pressing question is perhaps indeed—as Norman (2019, p. 197) suggests—how two *independent* heads compare with two *interacting* ones. We addressed precisely this question in a recent experimental study (Hautz et al. 2015), in which advanced medical students individually or in interacting teams of two diagnosed virtual patients presenting to the emergency room. Teams were more accurate than individuals (67.78 vs. 50.00%; difference 17.78% [95% CI, 5.83–29.73%]; $P = .004$), although knowledge levels were comparable and equal numbers of diagnostic tests were consulted before a diagnosis was made (Hautz et al. 2015). Most importantly for the current discussion, we found that interacting teams outperformed the same number of independently working individuals whose solutions were algorithmically aggregated. Specifically, we randomly paired students who had participated individually into simulated pairs (or "nominal groups") and used the diagnosis of the more confident member as this pair's diagnosis. We repeated this procedure 1000 times. The mean accuracy of simulated pairs (mean 56.73%; 95% CI, 49.72–63.74%) was comparable with that of individuals (mean 50%; 95% CI, 40.53–59.47%) but below that of real pairs (mean 67.78%; 95% CI, 59.95–75.6%; $F(2,83) = 6.75$, $\eta_p^2 = 0.12$; $P = .002$) (Hautz et al. 2015). Further simulations (Kämmer et al. 2017) showed that it would have taken aggregation of the opinions of three independent individuals according to a follow-the-most-confident rule [i.e., maximum confidence slating (Koriat 2012)] to outperform interacting pairs. Admittedly, this approach relies on the assumption that confidence is related to diagnostic accuracy on a per-case basis. In an analysis of a heterogeneous sample of 283 students working through the same cases individually, students were, on average, indeed more confident in diagnoses that were correct (mean confidence 57.3%; 95% CI, 54.2–60.3%) than in those that were not [41.8%, 95% CI, 39.1–44.6%; $F(1,253) = 196$; $P < .001$; $d = .63$)] (Hautz et al. 2019)—a result that is in line with other findings on self-monitoring (Eva and Regehr 2011; Pusic et al. 2015; Tweed et al. 2017).

In sum, evidence for the merits of aggregating independent heads (Kämmer et al. 2017; Kurvers et al. 2015; Surowiecki 2005) represents just part of the picture. A number of studies (Hautz et al. 2015; McMahon et al. 2016; Navajas et al. 2018) also provide evidence for interaction being beneficial to performance. Admittedly, these studies are still few in number, and more research is required into when and why interaction can outperform the wisdom of crowds.

How to conceptualize the problem?

In his editorial, Norman (2019) first discusses individual versus group-based *learning*, explicitly contrasting lectures to small group instruction. Although lectures are typically held in front of very large groups, they are often seen as individual instructions. This somewhat paradoxical classification results from the theoretical explanation, that the benefits of collaborative learning largely result from interaction with socially and cognitively congruent peers (Lockspeiser et al. 2008). Lectures are simply assumed to be too large to allow for such peer interaction. Norman later extends his argument to *performance* tasks such as diagnosis. However, the type of task heavily affects the answer (c.f., Soderstrom and Bjork 2015): learning may occur in solitude (e.g., through contemplation, observation, or reading) or in groups of various sizes. In learning groups, others merely constitute the environment that enables (or hinders) learning, which ultimately occurs in the individual. For learning tasks, it is thus possible to meaningfully compare individuals who studied

and perform alone with those who studied in groups but then perform alone. Several studies of motor skill learning conducted such comparisons and report —at least for complex skills—non-inferiority in effectiveness (and thus higher efficiency) of small-group versus individual learning (e.g., Räder et al. 2014; Tolsgaard et al. 2015).

In performance tasks, in contrast, the group is both the environment *and* the performing entity. With the possible exception of blatant individual errors, it is next to impossible to attribute the outcome of an interacting team's performance to any individual member of that team. Consequently, the performance of groups cannot be meaningfully compared with that of individuals in isolation in performance tasks [but we are guilty of reporting such comparisons ourselves (Hautz et al. 2015)]. Even if we do compare the performance of nominal groups of size *n* to that of interacting groups of the same size (rather than to the average or best individual), it remains an abstract comparison that offers only limited insights. Indeed, such a comparison might amount to comparing "treatment of myocardial infarction with aspirin and nitroglycerin to treatment with low-molecular weight heparin, primary angioplasty, beta-blockade, angiotensin-converting enzyme inhibitor, HMG-CoA-reductase inhibitor, clopidogrel, and folate—all at once. Even if a significant result is found in such an investigation, little is known about which therapies contributed" (Cook 2005, p. 542, referring to cross-media comparisons). Likewise, even if differences are found in the diagnostic performance of groups and individuals (whether alone or aggregated into nominal groups), it is impossible to say whether they are due to the number of people involved or result from specifics of the groups' configuration, behavior or the mode of aggregation. While nominal and interactive groups might both take advantage of an increasing amount of resources (such as knowledge, skills, cognitive capacity or experience) with increasing group size, interactive groups are affected by a number of phenomena that remain absent in nominal groups. On the downside, interactive groups have to coordinate their members, a demand that may result in process and motivation losses (Steiner 1972). In addition, mechanisms such as group conformity bias [the urge of a group member to conform with the stance of a majority (Kiesler and Kiesler 1969)], groupthink [group harmony taking priority over decision quality (Baron 2005; Janis 2013)], and polarization [groups taking more extreme decisions than any of their individual members is initially inclined to (Moscovici and Zavalloni 1969)] may reduce group performance even further. On the other hand, discussions among group members may yield new emergent solutions ("more than the sum of its parts") because of the combination of non-redundant information, for example (Kerr and Tindale 2004). Thus, even if the size and composition of nominal and interactive groups is kept equal, a comparison between the two—despite providing a necessary benchmark in many instances—can yield only vague insights into the origins of performance differences and is thus of limited use.

Why is it the wrong question?

Finally, we argue that the question of whether or not to work with colleagues on a particular problem is irrelevant in clinical practice. There, doctors either work in the seclusion of a private practice or are part of a clinical team (from the tumor board to ward rounds to multidisciplinary trauma teams). Although research may inform how such teams are formed and run, in most cases it would be difficult to substitute them by minds working in isolation (with or without later aggregation into a nominal group). The ever-changing demands of the health care environment will further increase the need for collaboration (Committee

on Diagnostic Error in Health Care et al. 2015). Driving factors include increased specialization, with no single individual having all knowledge relevant to a given patient's case, the accelerating rate of technological change, growing economic pressures resulting in the deregulation of health care, which necessitates collaboration across professions, and older patients with more complex conditions requiring multidisciplinary attention. All of these factors make individual practice increasingly difficult or even obsolete.

The rather black-and-white question of whether or not to collaborate to solve tough problems should nowadays be answered with a clear "it depends." As Norman (2019) notes, the idea of using brainstorming in interactive groups to solve virtually every problem (Osborn 1953) was later abandoned entirely in favor of individual decision making (Stroebe and Diehl 1994). But throwing out the baby with the bathwater in this manner was not the route to optimal performance either. The current recommendation is to combine individual and group phases of decision making, depending on the problem (Paulus and Kennworth 2019; Stroebe and Diehl 1994; Van De and Delbecq 1971). A combination or iterations of individual and group phases may likewise be suited for numerous healthcare problems. The diagnostic process, for example, comprises phases of information gathering, interpretation, and integration (Committee on Diagnostic Error in Health Care et al. 2015). Here, it may be advisable to consciously delegate specific phases to either groups or individuals. It remains an empirical question for future research which combination yields the best results for which task.

While most of this commentary is concerned with *performance* tasks, we suspect that our conclusion may also apply to individual versus group based *learning*. As Norman notes (2019), evidence for the effectiveness if small group instruction is mixed. Studies that focus on *knowledge* acquisition often reveal negative effects (Davis et al. 1992; Eagle et al. 1992). However, some studies that focus on the acquisition of *procedural skills* report rather positive consequences of small group instruction (e.g., Räder et al. 2014; Tolsgaard et al. 2015). Again, whether or not groups should be preferred over individuals seems to depend on the specifics of the task (such as learning verbal material or performing a motor task; see e.g. Soderstrom and Bjork 2015).

What are the consequences?

In a thorough discussion of the epistemological basis of medical education, Norman (2003) suggested that "the various factors that might contribute to a result [be] systematically varied over a series of experiments, based on a theory of causation, so that the real active ingredients in the treatment are understood" (p. 584). Decades of research, predominantly from sociology and social psychology, have provided plenty of theories on group performance that could and should be tested in their application to clinical environments. For example, a recent review concluded that the four factors information distribution, information exchange, heterogeneity in experience, and information retrieval from the team by its members all crucially affect performance in collaborative clinical reasoning (Kiesewetter et al. 2017). The role of these and other factors, however, is likely context depended. As just one example, consider team leadership. Teams diagnosing an unstable patient's condition in a trauma room benefit from directive leadership behavior, and "talking to the room" is associated with longer time to a correct diagnosis (Härgestam et al. 2016). In the context of ambiguous presentations in non-life-threatening situations, in contrast, collaborative leadership and talking to the room are associated with better performance (Tschan et al.

2009). In a review of the literature on how to lead trauma teams, Ford et al. (2016) emphasized the context dependency of leadership and concluded that "directive leadership is most effective when Injury Severity Score (ISS) is high or teams are inexperienced, while empowering leadership is most effective when ISS is low or teams more experienced" (Ford et al. 2016).

A thorough, theory-guided and methodologically sound investigation of when, why and how which type of teams (or individuals, for that matter) are best applied to which clinical problems may provide more valuable insights (Hautz et al. 2017) than a discussion about the general superiority of individuals, nominal or interacting groups.

References

Barnett, M. L., Boddupalli, D., Nundy, S., & Bates, D. W. (2019). Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA Network Open, 2*(3), e190096. https://doi.org/10.1001/jamanetworkopen.2019.0096.

Baron, R. S. (2005). So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. In James M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 37, pp. 219–253). Amsterdam: Elsevier. https://doi.org/10.1016/S0065-2601(05)37004-3.

Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, & The National Academies of Sciences, Engineering, and Medicine. (2015). Diagnostic team members and tasks: Improving patient engagement and health care professional education and training in diagnosis. In E. P. Balogh, B. T. Miller, & J. R. Ball (Eds.), *Improving diagnosis in health care*. Washington, DC: National Academies Press. https://doi.org/10.17226/21794.

Davis, W. K., Nairn, R., Paine, M. E., Anderson, R. M., & Oh, M. S. (1992). Effects of expert and non-expert facilitators on the small-group process and on student performance. *Academic Medicine, 67*(7), 470–474. https://doi.org/10.1097/00001888-199207000-00013.

Deloitte Insigths. (2019). 2019 Deloitte global human capital trends. *2019 Deloitte global human capital trends*. https://www2.deloitte.com/content/dam/insights/us/articles/5136_HC-Trends-2019/DI_HC-Trends-2019.pdf. Retrieved July 30, 2019.

Eagle, C. J., Harasym, P. H., & Mandin, H. (1992). Effects of tutors with case expertise on problem-based learning issues. *Academic Medicine, 67*(7), 465–469. https://doi.org/10.1097/00001888-199207000-00012.

Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education: Theory and Practice, 16*(3), 311–329. https://doi.org/10.1007/s10459-010-9263-2.

Ford, K., Menchine, M., Burner, E., Arora, S., Inaba, K., Demetriades, D., et al. (2016). Leadership and teamwork in trauma and resuscitation. *Western Journal of Emergency Medicine, 17*(5), 549–556. https://doi.org/10.5811/westjem.2016.7.29812.

Härgestam, M., Lindkvist, M., Jacobsson, M., Brulin, C., & Hultin, M. (2016). Trauma teams and time to early management during in situ trauma team training. *British Medical Journal Open, 6*(1), e009911.

Hautz, W. E., Kämmer, J. E., Exadaktylos, A., & Hautz, S. C. (2017). How thinking about groups is different from groupthink. *Medical Education, 51*(2), 229. https://doi.org/10.1111/medu.13137.

Hautz, W. E., Kammer, J. E., Schauber, 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.

Hautz, W. E., Schubert, S., Schauber, S. K., Kunina-Habenicht, O., Hautz, S. C., Kämmer, J. E., et al. (2019). Accuracy of self-monitoring: Does experience, ability or case difficulty matter? *Medical Education, 53*(7), 735–744. https://doi.org/10.1111/medu.13801.

Janis, I. L. (2013). *Groupthink: Psychological studies of policy decisions and fiascoes* (2 [Nachdr.] ed.). Boston: Wadsworth.

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.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55,* 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.

Kiesler, C. A., & Kiesler, S. B. (1969). *Conformity*. Reading: Addison-Wesley.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*(1), 80–113. https://doi.org/10.1037/a0025648.

Kurvers, R. H. J. M., Krause, J., Argenziano, G., Zalaudek, I., & Wolf, M. (2015). Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatology, 151*(12), 1346. https://doi.org/10.1001/jamadermatol.2015.3149.

Lockspeiser, T. M., O'Sullivan, P., Teherani, A., & Muller, J. (2008). Understanding the experience of being taught by peers: The value of social and cognitive congruence. *Advances in Health Sciences Education, 13*(3), 361–372. https://doi.org/10.1007/s10459-006-9049-8.

McMahon, K., Ruggeri, A., Kämmer, J. E., & Katsikopoulos, K. V. (2016). Beyond idea generation: The power of groups in developing ideas. *Creativity Research Journal, 28*(3), 247–257. https://doi.org/10.1080/10400419.2016.1195637.

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12*(2), 125–135.

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.

Norman, G. (2019). Two heads are better than one? *Advances in Health Sciences Education, 24*(2), 195–198. https://doi.org/10.1007/s10459-019-09888-3.

Osborn, A. (1953). *Applied imagination: Principles and procedures of creative thinking*. New York, NY: Charles Scribner's Sons.

Paulus, P., & Kennworth, J. (2019). Effective brainstorming. In P. Paulus & B. A. Nijstad (Eds.), *The oxford handbook of group creativity and innovation* (pp. 287–306). New York, NY: Oxford Library of Psychology.

Pusic, M. V., Chiaramonte, R., Gladding, S., Andrews, J. S., Pecaric, M. R., & Boutis, K. (2015). Accuracy of self-monitoring during learning of radiograph interpretation. *Medical Education, 49*(8), 838–846. https://doi.org/10.1111/medu.12774.

Räder, S. B. E. W., Henriksen, A.-H., Butrymovich, V., Sander, M., Jørgensen, E., Lönn, L., et al. (2014). A study of the effect of dyad practice versus that of individual practice on simulation-based complex skills learning and of students' perceptions of how and why dyad practice contributes to learning. *Academic Medicine, 89*(9), 1287–1294. https://doi.org/10.1097/ACM.0000000000000373.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000.

Stroebe, W., & Diehl, M. (1994). Why groups are less effective than their members: On productivity losses in idea-generating groups. *European Review of Social Psychology, 5*(1), 271–303. https://doi.org/10.1080/14792779543000084.

Surowiecki, J. (2005). *The wisdom of crowds* (1st ed.). New York, NY: Anchor Books.

Tolsgaard, M. G., Madsen, M. E., Ringsted, C., Oxlund, B. S., Oldenburg, A., Sorensen, J. L., et al. (2015). The effect of dyad versus individual simulation-based ultrasound training on skills transfer. *Medical Education, 49*(3), 286–295. https://doi.org/10.1111/medu.12624.

Tschan, F., Semmer, N. K., Gurtner, A., Bizzari, L., Spychiger, M., Breuer, M., et al. (2009). Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making. *Small Group Research, 40*(3), 271–300. https://doi.org/10.1177/1046496409332928.

Tweed, M., Purdie, G., & Wilkinson, T. (2017). Low performing students have insightfulness when they reflect-in-action. *Medical Education, 51*(3), 316–323. https://doi.org/10.1111/medu.13206.

Van De, A., & Delbecq, A. L. (1971). Nominal versus interacting group processes for committee decision-making effectiveness. *Academy of Management Journal, 14*(2), 203–212. https://doi.org/10.2307/255307.

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.