Critique and Rebuttal

Absence of DOA Effect but No Proper Test of the Lumberjack Effect: A Reply to Jamieson and Skraaning (2019)

Christopher D. Wickens, Alion Science and Technology, Boulder, CO, USA, **Linda Onnasch**, Humboldt-Universität zu Berlin, Germany, **Angelina Sebok**, TiER1 Performance, Denver, CO, USA, and **Dietrich Manzey**, Technical University Berlin, Germany

Objective: The aim was to evaluate the relevance of the critique offered by Jamieson and Skraaning (2019) regarding the applicability of the lumberjack effect of human–automation interaction to complex real-world settings.

Background: The lumberjack effect, based upon a meta-analysis, identifies the consequences of a higher degree of automation—to improve performance and reduce workload—when automation functions as intended, but to degrade performance more, as mediated by a loss of situation awareness (SA) when automation fails. Jamieson and Skraaning provide data from a process control scenario that they assert contradicts the effect.

Approach: We analyzed key aspects of their simulation, measures, and results which we argue limit the strength of their conclusion that the lumberjack effect is not applicable to complex real-world systems.

Results: Our analysis revealed limits in their inappropriate choice of automation, the lack of a routine performance measure, support for the lumberjack effect that was actually provided by subjective measures of the operators, an inappropriate assessment of SA, and a possible limitation of statistical power.

Conclusion: We regard these limitations as reasons to temper the strong conclusions drawn by the authors, of no applicability of the lumberjack effect to complex environments. Their findings should be used as an impetus for conducting further research on human–automation interaction in these domains.

Applications: The collective findings of both Jamieson and Skraaning and our study are applicable to system designers and users in deciding upon the appropriate level of automation to deploy.

Address correspondence to Christopher D. Wickens, Alion Science and Technology, Boulder, CO, USA; e-mail: pandawickens94@aol.com

The issue of performance consequences when humans interact with automated systems differing in their level or degree of automation (DOA) has become an important human factors topic stimulated by the seminal works of Sheridan and Verplank (1978) and Parasuraman et al. (2000). One general assumption that has become influential in this area is the so-called "lumberjack effect" or lumberjack model, introduced by Wickens and colleagues (e.g., Onnasch et al., 2014; Sebok & Wickens, 2017; see also Kaber, 2018). The assumption refers to effects of increasing DOAs on overall operator–system performance in (1) routine conditions, when the automation works reliably as intended, and (2) off-nominal or failure conditions, when the same automation fails or commits errors. Relevant operator–system performance parameters include operator situation awareness (SA) with respect to the functioning of the system, objective task performance, and perceived workload when using the system.

The analogy of the lumberjack effect ("the higher the trees are, the harder they fall") describes an assumed trade-off between higher DOAs and performance, in which overall task performance and workload are expected to benefit from higher DOAs as long as the automation works as expected (routine performance). However, these performance parameters are more adversely degraded in situations where the automation fails. Take as an example an

automated decision-support system (DSS) supporting the spatial orientation of surgeons during a noninvasive surgery (Manzey et al., 2011). In this case, the effect would predict that, as long as the system works reliably, surgeons' performance in terms of speed, accuracy, and workload would be better in the more highly automated DSS. However, in case the system fails or commits an error, higher DOA systems are expected to have more severe negative consequences on surgeons' performance than failures of less automated systems.

We have tested the hypothesized effect in a meta-analysis of a series of studies that were available from the literature at the time of the writing (2013) and which fulfilled several requirements. To be included in the analysis, studies were required to have varied the DOA of a task, assessed performance in the automated task, and include at least one of two other measures: performance on the same task when the automation unexpectedly failed and/or a measure of SA on the task, prior to the failure. A fourth measure, workload, was sometimes included if those data were provided by the authors of the study (Onnasch et al., 2014).

Based on a total of 18 studies we found support for the effect in terms of the predicted trade-off. Higher DOAs were associated with benefits in terms of increased routine task performance and reduced operators' workload. At the same time, however, decrements in failure performance and a loss of SA associated with increasing DOA were found, just as suggested by the lumberjack effect. However, an obvious drawback of this meta-analysis was the lack of studies with real operators in complex real-world settings, because not a great deal of such data were available at that time (2013). This was pointed out correctly by Jamieson and Skraaning (2018). As a consequence, one might challenge the extent to which the lumberjack effect is also valid in settings beyond the laboratory, although aviation accident and incident reports often point to these variables as potential causes (Airplane State Awareness Joint Safety Implementation Team, 2014).

Jamieson and Skraaning (2019) now have followed up on this question in carrying out a study which they argue to be a proper test of the lumberjack effect within a highly realistic scenario and employing well-trained professionals. We appreciate this effort as studies under realistic conditions are still an exception in human–automation research and are certainly highly needed. In particular, data of such studies will help researchers identify the extent to which the automation-induced performance consequences found in laboratory research and reflected in different variables are as strong, or weaker, or perhaps nonexistent in more complex and real user settings. Based on their own results and one other study (Calhoun et al., 2009) that was already identified in our meta-analyses as being in conflict with the lumberjack analogy, the authors draw strong conclusions that "our results fail to provide any support for the lumberjack model" (p. 12) and that "the lumberjack model has little to offer designers of complex human–machine systems" (p. 12).

We do not agree with such a strong statement because we believe that the design and metrics of the current study are not fully appropriate to permit such conclusions. We elaborate on this argument in more detail in the following.

1. *The nature of automation (whose degree is manipulated, and which fails)*. The authors based their data analysis on a human–automation interaction study in a nuclear power plant which originally was not planned as a test of the specific assumptions of the lumberjack model. Nevertheless, they identified the data as suitable for such test. The automated system targeted by the authors was the Computerized Operating Procedure System (COPS), representing an assistance system which provided control room operators with support in running test procedures of the underlying pressure relief system of the plant. COPS hosted the critical independent variable, *degree of automation* (DOA), and the authors describe its implementation in a manual condition and in three further DOAs support.

In order to test the lumberjack effect based on this system one would have expected an analysis of performance of the operators in situations where the system was available (routine performance) and when the system fails (failure performance). The latter would constitute situations where COPS suddenly is no longer available or where COPS makes errors, for example,

suggesting wrong steps to the operator or not informing the operator about a critical state of the pressure relief system (e.g., Wickens et al., 2015). However, it seems that what the authors term as "failure performance" (when a valve would fail to close) in fact describes a failure in the underlying pressure relief system, not the support system. Although this situation is interesting to inform how different DOA versions of COPS affect performance with respect to identifying and managing failures in the pressure relief system, it does not appear to present a fully proper situation for testing the main assumption of the lumberjack effect, in which the support system itself fails. Thus it may be that this paradigm difference creates the failure to replicate the two important negative trends in the lumberjack model (reduced failure performance and loss of SA). Indeed, it is certainly plausible that a higher DOA of a system that *assists* fault management (and does not itself fail) would itself *not* degrade fault management performance, and perhaps would actually improve that fault management. Jamieson and Skraaning indeed report that performance is not degraded when reporting their experimental results, as can be seen in their Figure 6.

2. *Situation awareness (SA) measure*. The authors place their greatest emphasis on their finding of an actual *improvement* of SA (with higher DOA), as assessed by their IPAQ measure. Anticipating that their measure might be criticized, they make an effort to defend it in the Discussion. Although we agree that prompting the operators' knowledge of the relative importance of certain parameters for a given system state might assess one aspect of SA, it nevertheless has critical limitations. First, it seems to be at least partially confounded with general system knowledge. Second, while it is important to conclude, as the IPAQ measure reveals, that operators know the relative importance of different parameters, this measure of general system knowledge is very different from the measure of awareness of the dynamic changing *value* (and not just importance) of specific process parameters during the evolution of the failure. It is knowledge of this value, which the authors did not assess, that is revealed by SA measures such as SAGAT (Endsley, 1995, 2000). An operator

can know at a particular point in time the *importance* of knowing the state of a valve; but that is very different from knowing *whether* that state is "open" or "closed" or whether an entire procedure has been executed as intended and, if not, what the deviations were. Thus, while we acknowledge the importance of this new measure of operator meta-cognition (IPAQ), we would argue that it is confounded with basic system knowledge and may not have assessed the most important aspect of SA which would underlie differences in failure response. This aspect would be needed for a proper test of the lumberjack model.

3. *Understating the subjective ratings of out-of-the-loop performance*. In order to get a subjective view of the operators participating in the study on the possible performance consequences of the different variants of COPS, participants were asked to evaluate different aspects related to what is referred to as "out-of-the-loop" performance issues in human–automation interaction (Endsley & Kiris, 1995). These questions directly prompt main assumptions of the lumberjack effect. These results clearly provide a statistical reliable trend that is *consistent* with the model. However, the authors discount these data with reference to their null effects in performance measures. This seems to be related to a bias in down-weighting data which do not fit the general argument. While subjective ratings admittedly can be challenged with respect to validity, we argue that the results of the subjective ratings should have been given more weight in this particular study, given that the ratings focus most closely on the relevant research question. More specifically, it is not easy to understand why such information "has little to offer designers" (p. 12). If this is the case, the question might be raised as to why subjective data were collected in the first place.

4. *Possible lack of statistical power*. The authors report eight crews in each counter-balanced condition of the repeated measures design. Depending on the amount of variance between crews, it is possible that there is insufficient statistical power to reveal the differences in failure performance, shown to be nonsignificant in their Figure 6. When making critical claims regarding a proof of the null hypothesis (e.g., a

null effect), it is incumbent on the researchers to provide sufficient statistical power to establish that effect if it were present, and to report that power.

5. *The lack of a comparison between performance in routine and automation failure conditions.* The lumberjack effect compares performance in normal and off-nominal situations and hypothesizes that performance will be improved by automation in routine conditions, but will be more adversely affected in automation failure conditions. The authors explicitly state "No measure of routine task performance was included in the study" (p. 7). Without such a comparison, this study does not fully evaluate the lumberjack effect.

In summary the authors assert they have found evidence, from their study and their assessment of a few other studies conducted in realistic work settings, that contradicts the lumberjack effect and results of a meta-analysis. From our assessment this conclusion is based on a study whose experimental design does not seem to provide a proper test of the model, which includes an SA measure that we believe does not capture awareness of the *current state* of the system parameters, which has limited statistical power, and which has reported a change in subjective ratings of performance *consistent* with that lumberjack effect. Furthermore, the assessment of other studies seems to be somewhat biased as (methodological) caveats were placed on studies that support aspects of the lumberjack model, yet no such caveats were placed on contradictory findings.

While we admire the authors' efforts to push predictions of the lumberjack model into a realistic scenario, we do not consider the study and the presented analysis of the literature as being an appropriate test of the hypothesized effects predicted by the model. Thus, the authors' more extreme overall claims that "the lumberjack model has little to offer designers of complex human–machine systems" (p. 12) and that "results fail to provide any support for the lumberjack model" (p. 12) do not seem to be warranted, based on the evidence presented. Thus, the applicability of the lumberjack model to complex human–automation systems in real work settings remains an open issue for further research. We encourage others to join Jamieson and Skraaning in pursuing this research.

KEY POINTS

- Jamieson and Skraaning (2019) have criticized the lumberjack effect of human–automation interaction as not applicable to real world operators in a complex process control simulation, on the basis of their findings that fail to replicate key features of the effect, particularly the loss of SA with a higher DOA.
- We present the counterargument that their critique is not entirely valid because their implementation of the automation that fails and their assessment of SA are not valid, along with other concerns.
- Evidence of the extent to which the lumberjack effect scales up to complex simulations remains ambiguous and more research of the sort conducted by Jamieson and Skraaning is required before a strong conclusion of nonapplicability, such as they state, can be offered with confidence.

ORCID iD

Christopher D. Wickens ⓘ https://orcid.org/0000-0002-6789-4446

REFERENCES

Airplane State Awareness Joint Safety Implementation Team. (2014). *Airplane state awareness joint safety implementation team final report*. https://www.cast-safety.org/pdf/JSAT-ASA_FinalReport_June2014.pdf.

Calhoun, G. L., Draper, M. K., & Ruff, H. A. (2009). *Effect of level of automation on unmanned aerial vehicle routing task* [Conference session]. Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting, Santa Monica, CA, Human Factors and Ergonomics Society, 197–201.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37, 65–84.

Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 131–157). Lawrence Erlbaum.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37, 381–394.

Jamieson, G. A., & Skraaning Jr., G. (2018). Levels of automation in human factors models for automation design: Why we might consider throwing the baby out with the bathwater. *Journal of Cognitive Engineering and Decision Making*, 12, 42–49.

Jamieson, G. A., & Skraaning Jr., G. (2019). The absence of degree of automation trade-offs in complex work settings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872081984270. online first.

Kaber, D. B. (2018). Issues in human–automation interaction modeling: Presumptive aspects of frameworks of types and

levels of automation. *Journal of Cognitive Engineering and Decision Making*, *12*, 7–24.

Manzey, D., Luz, M., Mueller, S., Dietz, A., Meixensberger, J., & Strauss, G. (2011). Automation in surgery: The impact of navigated-control assistance on performance, workload, situation awareness, and acquisition of surgical skills. *Human Factors*, *53*, 584–599.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*, 476–488.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*, 286–297.

Sebok, A., & Wickens, C. D. (2017). Implementing lumberjacks and black swans into model-based tools to support human–automation interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*, 189–203.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). MIT, Man Machine Systems Laboratory.

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*, 728–739.

Christopher D. Wickens is a professor emeritus of aviation and psychology at the University of Illinois and is a senior scientist at Alion Science and Technology, Boulder, CO, and professor of psychology at Colorado State University. He received his PhD in psychology from University of Michigan in 1974 and served three years in the U.S. Navy.

Linda Onnasch is assistant professor of engineering psychology at the Humboldt-Universität zu Berlin, Germany. She earned her PhD in psychology from the Technische Universität Berlin, Germany, in 2015.

Angelina Sebok is a senior research scientist at TiER1 Performance. She earned her MS in industrial and systems engineering from Virginia Tech and is a Certified Human Factors Professional.

Dietrich Manzey is professor of work, engineering and organizational psychology in the Department of Psychology and Ergonomics at the Technische Universität Berlin, Germany. He earned his PhD in psychology from the University Kiel, Germany, in 1988 and got his habilitation from the University of Marburg, Germany, in 1999.