JUST RELEASED!

Workload Assessment: A Guide for Researchers, Practitioners, and Students

Volume 2, Users' Guides to Human Factors and Ergonomics Methods

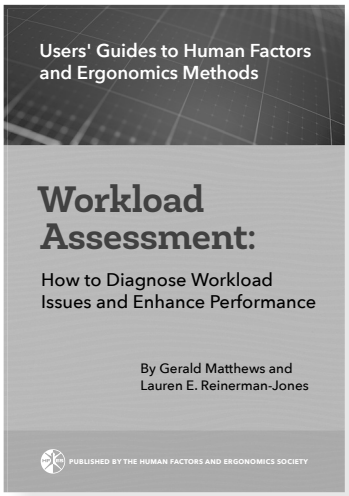
Workload assessment is important wherever people perform under high levels of task demand, such as multitasking, time pressure, and interacting with complex interfaces. This accessible guide sets out a comprehensive, systematic approach to choosing and evaluating workload measures and to designing studies to maximize the value obtained from the measures.

No other single volume in the current literature deals exclusively with workload assessments. In this book, you'll find

- **Basic concepts in both workload theory and applications in a variety of domains**
- **A comprehensive survey of leading self-report, performance-based, and psychophysiological measures**
- **A checklist to ensure assessment quality**
- **Two detailed workload examples to illustrate practical applications.**

Workload Assessment has been written to be accessible to a wide audience and generally requires little specific background knowledge. This book will help guide **researchers** toward best practices in the use of workload measures to test theory-driven hypotheses in studies of cognitive psychology and cognitive neuroscience. **Practitioners** in domains such as surface transportation, aerospace, industrial ergonomics, the military, cybersecurity, system design, education, and health care will be able to choose the most appropriate workload measures for applied problems, and use workload data in efforts to mitigate performance issues. *Workload Assessment* is essential reading for graduate **students** in human factors and applied cognitive psychology, as well as supplementary reading for undergraduate students in these topics.

ISBN 978-0-945289-51-7
130 pp., 7" x 10" paperback and e-book
http://www.hfes.org/publications/



Gerald Matthews, PhD (University of Cambridge), is a research professor in the Institute for Simulation and Training, University of Central Florida. He previously held faculty positions at the University of Cincinnati, University of Dundee, and Aston University. His research centers on various human performance issues, including workload, stress, fatigue, and individual differences factors.

Lauren Reinerman-Jones, PhD (University of Cincinnati), is director of Prodigy, a lab at the University of Central Florida's Institute for Simulation and Training. Her research focuses on assessment for explaining, predicting, and improving human performance and system design in a variety of domains including nuclear, human-robot teaming, training and education, medical, aviation, and cyber.

PUBLISHED BY THE HUMAN FACTORS AND ERGONOMICS SOCIETY

V 62 | N 6 | SEPTEMBER 2020

HUMAN FACTORS
The Journal of the Human Factors and Ergonomics Society

CONTENTS

■ HEALTH CARE/HEALTH SYSTEMS

954 An Experimental Validation of Masking in IEC 60601-1-8:2006-Compliant Alarm Sounds
Matthew L. Bolton, Xi Zheng, Meng Li, Judy Reed Edworthy, and Andrew D. Boyd

This research used signal detection experiments to validate that the psychoacoustics of simultaneous masking could predict when IEC 60601-1-8-compliant medical alarm sounds are audible based on the masking of alarm primary harmonics. The results will inform the IEC 60601-1-8 standard and methods for detecting masking in medical alarms.

■ HUMAN-COMPUTER INTERACTION, COMPUTER SYSTEMS

973 Classification of Attentional Tunneling Through Behavioral Indices
Sean W. Kortschot and Greg A. Jamieson

Operators managing information dense systems are faced with heavy attentional demands, which can result in attentional tunnels. Behavioral rather than physiological correlates can be used to infer when operators are attentionally tunneled. This research represents a proof of concept for using operator behavior to infer their attentional state.

987 Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming
Na Du, Kevin Y. Huang, and X. Jessie Yang

We summarized existing literature and examined the effects of displaying three types of likelihood information on human-automation team performance using a simulated surveillance task. Results indicate presenting hits and correct rejection rates alone should be avoided.

■ MOTOR BEHAVIOR

1002 Effects of Initial Starting Distance and Gap Characteristics on Children's and Young Adults' Velocity Regulation When Intercepting Moving Gaps
Hyun Chae Chung, Gyoojae Choi, and Muhammad Azam

We examined young adults' and children's velocity regulation in gap crossing. We manipulated initial starting distance and gap characteristics, including inter-vehicle gap and vehicle size. Children did not finely tune their movements when they approached large moving vehicles from closer distances.

■ SURFACE TRANSPORTATION

1019 Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures
Anthony D. McDonald, Thomas K. Ferris, and Tyler A. Wiener

This study compares methods for classifying cognitive distraction and texting from normal driving. Positive directions for future work are identified through comparisons of a broad set of driver-based and physiological input measures, a comprehensive feature generation process, and several common machine learning approaches.

HUMAN FACTORS

The Journal of the Human Factors and Ergonomics Society

Human Factors: The Journal of the Human Factors and Ergonomics Society (HFES) publishes original articles about people in relation to systems, including hardware, software, and environmental factors. Evaluative reviews of the literature, definitive articles on methodology and procedures (including prescriptive methods for task simulation and evaluation), quantitative and qualitative approaches to theory (including models of human performance and safety), and empirical articles reporting original research are considered for publication. The opinions and judgments expressed in this journal are those of the authors, and are not necessarily those of the editor; neither are they to be construed as representing the official policy of the Human Factors and Ergonomics Society.

Human Factors also publishes special sections that focus on important areas of human factors/ergonomics in an integrated manner. *Human Factors* is endorsed by the International Ergonomics Association (http://www.iea.cc).

Manuscripts should be submitted to *Human Factors* via the following Web site: http://mc.manuscriptcentral.com/humanfactors. All submissions to *Human Factors* must be uploaded to the journal site for consideration. Please visit http://hf.sagepub.com/ to view submission guidelines. Contact the editorial review coordinator (journal@hfes.org) if you encounter difficulties with your online submission.

Human Factors: The Journal of the Human Factors and Ergonomics Society (ISSN 0018-7208) (J632) is published eight times annually—in February, March, May, June, August, September, November, and December—by SAGE Publishing, 2455 Teller Road, Thousand Oaks, CA 91320 on behalf of the Human Factors and Ergonomics Society, 2025 M Street NW, Suite 800, Washington, DC 20036 USA. Periodicals postage paid at Thousand Oaks, California, and at additional mailing offices. POSTMASTER: Send address changes to the Human Factors and Ergonomics Society, 2025 M Street NW, Suite 800, Washington, DC 20036 USA.

Subscription Information: All subscription inquiries, orders, back issues, claims, and renewals should be addressed to SAGE Publishing, 2455 Teller Road, Thousand Oaks, CA 91320; telephone: (800) 818-SAGE (7243) and (805) 499-0721; fax: (805) 375-1700; e-mail: journals@sagepub.com; Web site: http://journals.sagepub.com. **Subscription Price:** Institutions: $1,115; Individuals: $589. For all customers outside the Americas, please visit http://www.sagepub.co.uk/customerCare.nav for information. **Claims:** Claims for undelivered copies must be made no later than six months following the month of publication. The publisher will supply replacement issues when losses have been sustained in transit and when the reserve stock will permit.

Member Subscription Information: Human Factors and Ergonomics Society member inquiries, changes of address, back issues, claims, and membership renewal requests should be addressed to Human Factors and Ergonomics Society, 2025 M Street NW, Suite 800, Washington, DC 20036 USA; telephone: (202) 367-1114; fax: (202) 367-2114; Web site: http://www.hfes.org; e-mail: membership@hfes.org. Requests for replacement issues should be made within six months of the missing or damaged issue. Beyond six months and at the request of HFES, the publisher will supply replacement issues when losses have been sustained in transit and when the reserve stock permits. Members also receive complimentary subscriptions to the annual *Directory & Yearbook*, the monthly *HFES Bulletin*, and the quarterly journals *Ergonomics in Design* and *Journal of Cognitive Engineering and Decision Making*.

Copyright Permission: To request permission for republishing, reproducing, or distributing material from this journal, please visit the desired article on the SAGE Journals website (journals.sagepub.com) and click "Permissions." For additional information, please see www.sagepub.com/journalspermissions.nav.

Change of Address for Nonmembers: Six weeks' advance notice must be given when notifying of change of address. Please send the old address label along with the new address to the SAGE office address above to ensure proper identification. Please specify name of journal, *Human Factors*.

Printed on acid-free paper

At the Forefront of HF/E

A Fundamental Cognitive Taxonomy for Cognition Aids

Anne Collins McLaughlin, North Carolina State University, Raleigh, USA, and **Vicky E. Byrne**, KBR, Houston, TX, USA

Objective: This study aimed to organize the literature on cognitive aids to allow comparison of findings across studies and link the applied work of aid development to psychological constructs and theories of cognition.

Background: Numerous taxonomies have been developed, all of which label cognitive aids via their surface characteristics. This complicates integration of the literature, as a type of aid, such as a checklist, can provide many different forms of support (cf. prospective memory for steps and decision support for alternative diagnoses).

Method: In this synthesis of the literature, we address the disparate findings and organize them at their most basic level: Which cognitive processes does the aid *need* to support? Which processes *do* they support? Such processes include attention, perception, decision making, memory, and declarative knowledge.

Results: Cognitive aids can be classified into the processes they support. Some studies focused on how an aid supports the cognitive processes demanded by the task (aid function). Other studies focused on supporting the processes needed to utilize the aid (aid usability).

Conclusion: Classifying cognitive aids according to the processes they support allows comparison across studies in the literature and a formalized way of planning the design of new cognitive aids. Once the literature is organized, theory-based guidelines and applied examples can be used by cognitive aid researchers and designers.

Application: Aids can be designed according to the cognitive processes they need to support. Designers can be clear about their focus, either examining how to support specific cognitive processes or improving the usability of the aid.

Keywords: cognitive aids, cognitive psychology, attention, memory, checklists, crisis checklists

Address correspondence to Anne Collins McLaughlin, Department of Psychology, North Carolina State University, Box 7650, Raleigh, NC 27695, USA; e-mail: anne_mclaughlin@ncsu.edu

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 865–873
DOI:10.1177/0018720820920099
Article reuse guidelines: sagepub.com/journals-permissions

At the core of human factors psychology is the study of human capability and limitation, particularly the limitations of human cognition. This includes our limited memories, both working and long term, our limited attention, and our biases in judgment and reasoning, to name a few. Our ability to achieve as a species is partly due to our use of tools that allow us to overcome these limitations, from a recipe that turns the science of baking into a series of prompted steps to the checklists used by pilots in emergencies to the formal analysis of competing hypotheses that intelligence analysts use to prevent biased decisions. These tools can broadly be considered "cognitive aids," in that they provide support for the cognitive processes likely to fail during a complex or demanding task. The term cognitive aid first appeared in the 1970s (e.g., Hormann, 1971) and was often used to describe decision-support systems (e.g., Aldag & Power, 1986) before being used more generally to describe systems that supported other cognitive processes (e.g., Reason, 1987). However, results are mixed in the literature, with some reviews finding little or no benefit to introducing an aid (Bosk et al., 2009; Marshall, 2013) and others finding great benefit, hindered only by lack of aid adoption (Chaparro et al., 2019).

It is possible some of the mixed findings have been due to a focus on the form of the aid rather than its function. For example, cognitive aids have taken the form of checklists, maps, sensors, and automation (see Hales et al., 2008; Marshall, 2013; Myers, 2016, for various reviews). All of these formats could and do support a variety of cognitive processes, from different types of memory to visual attention, divided attention, and so on. Designing or comparing aids according to format misses the larger picture of the function of an aid: to support cognitive processes. Burian and colleagues suggested cognition must be

considered to understand "why clinical checklists are or are not effective in different settings" (Burian et al., 2018). It is not surprising clinical checklists would not be effective if their function were to support the wrong cognitive process. For example, the original study of the WHO Surgical Safety Checklist noted that the step requiring patient and surgery site identification was a new procedure for some of the hospitals in their study (Haynes et al., 2009). For those sites, the checklist functioned to increase knowledge. At sites already using the procedure, it functioned to focus attention on the attributes of the surgical case. If the step were worded to support only the focus of attention by experts, and assumed the knowledge of how to check the identification of patient and site was present, those needing a knowledge aid might not benefit from the checklist. Thus, we agree with Burian and colleagues and extend the question beyond checklists to all cognitive aids. In this article, we propose to organize aids at their highest level: the type of cognition the aid should support. By doing so, we believe that new aids can be chosen or designed using well-established psychological theory and that the domain of cognitive aids will inform these theories. Essentially, a taxonomy of design choices can be created and tested to match the desired function of any cognitive aid.

REVIEW OF PREVIOUS TAXONOMIES

There have been efforts to classify cognitive aids according to useful categories, such as call-and-response checklists, "job aids," or medical cognitive aids (Hales et al., 2008; Marshall, 2013; Winters et al., 2009). These taxonomies were restricted to a domain (e.g., medicine or aviation) with implicit assumptions of the kinds of tasks prevalent in that domain. Indeed, important differences exist between application domains (Kapur et al., 2015). Context was not explicitly discussed in these taxonomies, perhaps because it was implicitly understood. For example, a taxonomy of checklists for the aviation domain considered "functions, format, design, length, usage, and the limitations of the humans who must interact with it" but not context or task type as a differentiator. However, it was likely assumed that the tasks were complex procedures in aviation and that the operators were already highly trained, making a posttask call-and-response checklist a beneficial option (Degani & Wiener, 1993). The posttask call-and-response checklist would not be suitable for other tasks, such as aiding first-time users of a cardiac defibrillator.

No taxonomy specified the cognitive processes being aided. Burian et al. (2018) made an important point concerning these taxonomies that the variety of cognitive aids that are called "checklists" have such varying attributes to make the term nondescriptive. For example, one checklist historically categorized as "medical" was not—it actually assisted doctors with business tax estimation (DeMuth & Achorn, 1978). Other medical checklists assisted teams during surgical procedures (World Health Organization [WHO], 2015) or provided decision support during cardiac arrests (Field et al., 2014). This is one reason the term cognitive aid is preferred—checklist denotes format, not function. Burian et al. (2018) presented a second taxonomy for medical checklists that characterized them by the types of cognition and actions they supported. Types of aids, such as critical event checklists or screening instruments, could be classified by their strengths: Were they designed more to aid memory or facilitate decision making? Were they needed in real time or after the fact? This was the only taxonomy to use cognitive processes and task demands as a classifier.

In sum, the previous taxonomies of cognitive aids focused on a particular field (e.g., medicine), often on a particular format (e.g., the checklist), and a particular focus (e.g., task type). This has practical value and offers a way to view the literature on cognitive aids. However, it is difficult to come to broader conclusions about the design of cognitive aids, particularly if the task to be aided is novel or differs from tasks with known successful aid designs. To develop a taxonomy informed by cognitive theories, we examined empirical articles investigating the effects of cognitive aids from 1998 to 2018 (literature review selection method detailed in McLaughlin & Byrne, 2019). What we found missing from these articles was consideration of the cognitive processes that are being aided.

Support of the Task by Cognitive Aid

The literature on cognitive aids can be divided into two areas of study: development and use of cognitive aids to help operators achieve an outcome (e.g., Kim & Dey, 2009) and studies on the usability of the aids (e.g., Clebone et al., 2019). In the literature, attributes of aid usability have been called contextual factors (Marmaras & Kontogiannis, 2001) and "event design elements" (Hepner et al., 2017). Without dividing the literature into studies of function versus usability, it is impossible to connect the cognitive process that the aid supports to outcome. For example, increasing the font size of a checklist may ease the perceptual demands of the operator when using the aid, but it would not ease the perceptual demands of the task for which the checklist was designed. We could not make conclusions about perception aids from such a study, though we could make usability conclusions regarding checklists. Unfortunately, the literature has not adopted this categorization. Figure 1 shows a sampling of the potential cognitive supports an aid may supply to an operator for achieving an outcome or making the cognitive aid easier to use. The processes that could be aided are known as limited cognitive resources: attention (Kahneman, 1973; Wickens, 2002), memory (Atkinson & Shiffrin, 1968), and perception (e.g., vision: Gibb et al., 2008; vestibular senses: Shappell et al., 2007). Also included were higher order processes affected by limited cognitive resources, such as decision making (De Martino et al., 2006) and learning/declarative knowledge (Stern, 2017). These processes form a theory-based start for the classification of aids, though there may be other processes that need to be added.

How Cognitive Processes Are Assisted by Cognitive Aids

Figure 1 lists a number of cognitive processes an aid could support. The list was meant to provide a starting point for categorizing the literature on aids by assessing which of these processes they were intended to support. Here, we provide examples of attentional aids from the literature. Gaps in the literature are discussed.

Attentional aids. It is generally accepted that attention is a limited resource composed of several fairly independent components (Navon & Gopher, 1979; Wickens, 2002). When a task or cognitive aid requires more resources than an operator possesses, performance declines. We provide an overview of aids supporting selective attention and orienting, divided attention, and sustained attention.

Selective attention is the ability to focus attention on important stimuli while ignoring others. This may mean to filter out distracting information (Broadbent, 1958) or that the distracting information is attenuated to the degree that it no longer consumes attentional resources (Treisman, 1964). Any aid supporting this cognitive process succeeds when it either helps to focus the operator on the important stimuli for the task or helps to inhibit the distracting stimuli in the task. Attentional *orienting* is a similar process, as attention is focused on a particular stimulus, with the small difference that there is the implication that an external cue caused the response, such as a flashing light to attract attention to a target (Müller & Rabbitt, 1989).

An aid can directly support noticing of cues in the task itself, leading to selective attention. For example, an aid might direct an operator to a cue needed for decision making as in the diagnosis of tension pneumothorax via the Stanford Anesthesia Aid: "Unilateral breath sounds, possible distended neck veins and deviated trachea (late signs)" (Howard et al., 2013). This directs attention toward cues that were potentially already present in the environment, but not noticed by the anesthesiologist. Many "job-aids" also support noticing in a task. For example, all warning systems and alarms direct attention, though memory recall is often required to interpret their meaning. In a European Space Agency project, augmented reality was implemented to aid noticing of warning signs and activity symbols while completing an ISS procedure for installing a temporary stowage rack (Helin et al., 2018). Most "crisis checklists" are also designed to support selective attention in the task (e.g., Goldhaber-Fiebert & Howard, 2013).

Other attentional cues can be provided within the aid to call attention to the structure, organization, or hierarchy of the aid. This supports usability of the aid. An example of attentional

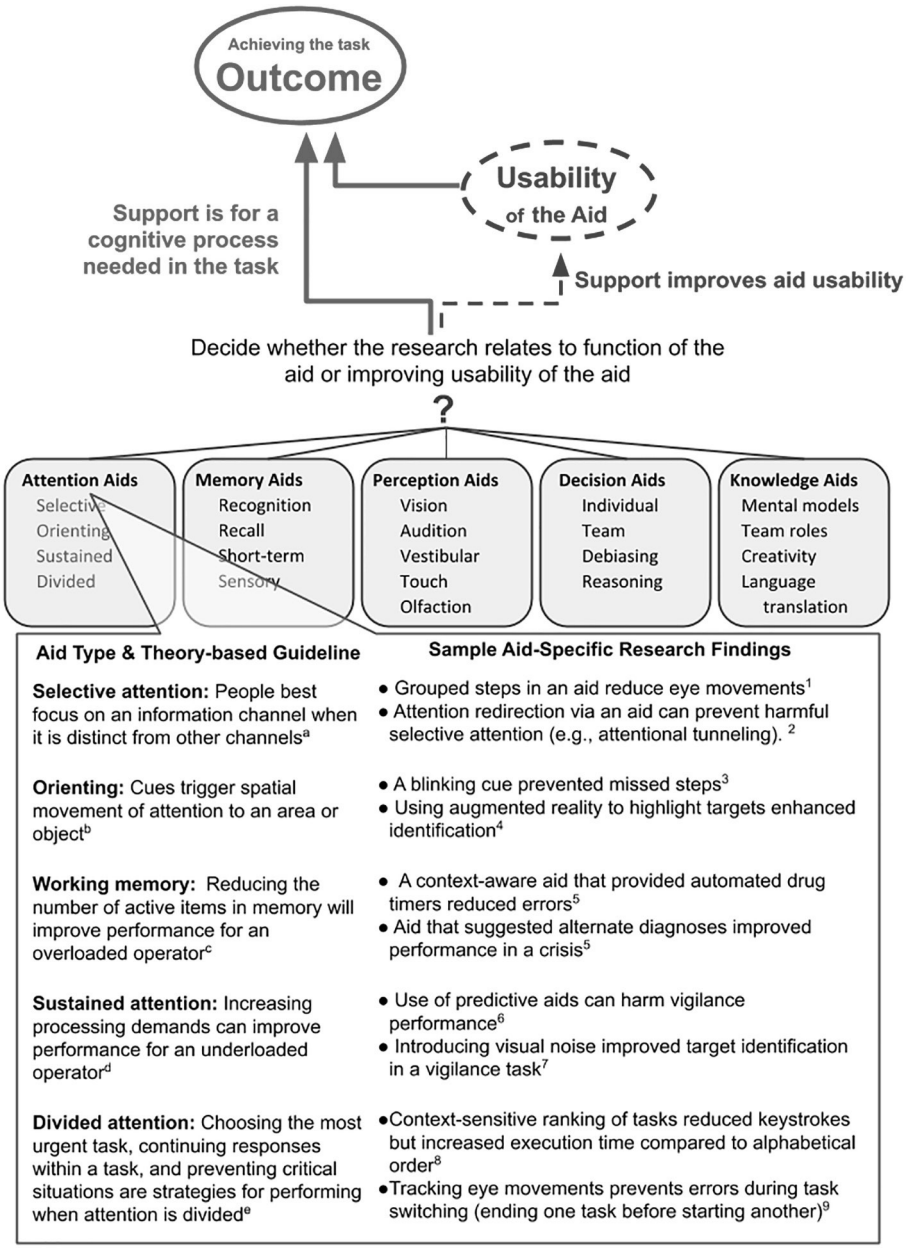


Figure 1. Model of a way to organize the literature on cognitive aids. Arrows show the potential support paths. Some support the task itself, such as supporting working memory by introducing an aid that collects and presents only the information necessary for a decision. Other supports may make the aid itself easier to use, such as increasing the contrast or font size of the aid interface. A detailed explanation of how the literature on aids could be presented is shown for attention aids. The same treatment may be given to the other cognitive processes that aids can support. [a]Treisman (1964); [b]Folk and Gibson (2001); [c]Wickens, Hollands et al. (2015); [d]Healy and Bourne (2013); [e]Rill et al. (2018); [1]Clebone et al. (2019); [2]Long et al. (2017); [3]Chung and Byrne (2008); Rusch et al. (2013); [5]Wu et al. (2014); [6]Minotra and McNeese (2017); Bodala et al. (2016); [8]Ramachandran et al. (2014); [9]Ratwani and Trafton (2011).

capture through usability improvements can be found in the WHO Surgical Safety Checklist (Haynes et al., 2009). The design of the checklist includes color-blocked backgrounds that divide the checklist into sections. Inside these sections, broad horizontal lines delineate subsections. Bold font denotes roles and actions. It is expected that attention moves back and forth between the checklist and the task, with these attentional cues aiding in visual reorientation to the aid. Wu et al. (2014) compared standard text, structured text, color block, pictographs, and interactive focus with expert users: eye tracking revealed how the color block design drew the user's gaze to the block that contained the information, even when looking back and forth from the primary task.

Divided attention means to distribute attention to *n* simultaneous tasks. Models of task switching help predict how an operator will respond to multiple task demands (Barg-Walkow & Rogers, 2017; Wickens & Gutzwiller, 2017) and this knowledge can be used in cognitive aid design. Such models suggest the importance of context-sensitive aid design, where notifications and alarm salience are adjusted according to the current importance of a task (e.g., Abdi et al., 2015). In the most notorious aviation example, an alarm regarding an interior light malfunction overcame the salience of a terrain warning system, resulting in a crash (Wiener, 1977). Eye tracking can also be used to detect when a task is or will be neglected, allowing an aid to prompt the return of attention to that task (Ratwani & Trafton, 2011). However, there is still much to be learned and implemented regarding aids for divided attention: Alarm fatigue (e.g., Kobayashi et al., 2017) and ineffective multitasking (Ralph et al., 2014) are still common issues.

Sustained attention, also called vigilance, is the process of maintaining alertness over periods of time (Davies & Parasuraman, 1982). Aiding sustained attention often means to reduce the need for it, via alarms (e.g., Fletcher & Bedwell, 2014), or other ways to augment slowly changing or rare stimuli that might be overlooked (Rall et al., 2010). Ironically, aids that support operators well can actually put more load on the operator's sustained attention resources. When cognitive aids predict events or outcomes, the operator may come to depend on that prediction and miss critical events not predicted by the aid (Minotra & McNeese, 2017). This is similar to the paradox of automation—when highly automated and reliable systems fail, it is more likely the failure will go unnoticed than in less automated or reliable systems (e.g., Bainbridge, 1983).

Another form of vigilance aid would be a *cognitive antidote* (Healy & Bourne, 2013): an increase in task difficulty or the addition of another task to keep the operator engaged. However, the limits, effectiveness, and possible detrimental effects of cognitive antidotes have not been studied regarding cognitive aids. There was little in the literature regarding cognitive aids for vigilance tasks with the exception of a study introducing visual noise in a virtual world (e.g., heavy rainfall) as an aid to maintaining vigilance for longer time periods (Bodala et al., 2016). The lack of research on cognitive antidotes is brought to light when an aid taxonomy highlights known cognitive processes and the (lack) of accompanying research.

To support *working memory* means to support attentional control and manipulating information in memory. When an aid directly supports the task via support of working memory, this usually means that the aid holds the information to be used in the task in an easily accessible format, so that information can be integrated and used by the operator. Most of the healthcare aids, particularly "handoff" aids for transitioning patients between care teams, supported the working memory of the operators by calling for the information crucial to clinical decisions and organizing that information for the operator (e.g., Weiss et al., 2013). Categorizing these aids as working memory aids was done post hoc, as studies do not currently report specific ways aids support cognition.

Case Study of Applying the Taxonomy

Miller et al. (2000) introduced three different cognitive aids designed to improve aircraft identification by military radar operators who had difficulty knowing whether they had previously identified an aircraft. Here, we retrospectively

discuss one of those aids in terms of the cognitive processes it supported. Aids were created to play a tone any time a new aircraft entered the display, use colors to mark an airplane as identified, and provide a symbol to indicate which aircraft needed to be identified quickly. The tone and colors supported *selective attention* by indicating which aircraft were already identified; the operator could selectively attend only to the unidentified aircraft. The tone and symbol for important targets supported *orienting* by drawing attention to specific targets. The combined cues supported *working memory* by placing the variables to be considered in aircraft identification "in the world" rather than relying on them being kept active "in the head" of the operator (Norman, 2013). Choosing color and sound as the methods of support was the usability consideration for the use of the aid.

DISCUSSION

The terms checklist, procedure, and cognitive aid have little meaning on their own, even when broken down into subtypes, such as "call and response" or "emergency procedure" or "decision aid." This is because any task that requires an aid is likely complicated enough to necessitate multiple forms of cognition support. We provide a general structure of categorization, where any single aid is broken down into how it provides support for multiple processes. A surgical checklist, for example, might provide information support via one step where all team members must announce their name and role, but in other steps, it might provide decision support for actions to take in event of heart or lung stoppage (McLaughlin et al., 2016). Another type of aid, role cards attached to surgical team members, would provide the same information support as the aforementioned checklist could, but is not a checklist (Renna et al., 2016). A different checklist might provide none of those supports (WHO, 2015). We propose comparing aids via the cognitive supports they provide. This will allow comparison across aids, whether those aids be checklists, flowcharts, static or dynamic, or any of the other labels used to describe cognitive aids. Aid researchers and designers can turn to the literature for guidelines on supporting the cognitive processes in their task of interest and see specific design choices others have attempted and their ramifications.

Application

Our method of classification is similar to (and depends on) task analysis. A complete understanding of the outcome desired and the demands of the task preceding that outcome is the crucial first step toward developing new cognitive aids. With a complete task analysis, task demands can be matched to the kind of cognitive support required. Using previously classified aids, types of previously successful support can be nominated for use in a new aid. For example, if the task has prospective memory demands, the aid should support prospective memory using evidence-based designs: alarms, notifications, or checklist steps. An organized literature will allow practitioners to consider previous aid designs during their design process.

Though such classification of aids will organize the literature and allow easier comparisons of studies and the development of new aids, it is not the only important piece of the picture. As previous taxonomies discussed, the designers of any aid and its components must consider the support options in terms of the functional task environment: "the moment-to-moment intersection of a cognitive agent (human operator) pursuing a particular goal in a particular physical environment" (Gray et al., 2006, p. 101). As mentioned by Hales et al. (2008), aid design interacts with the limitations of physical storage and use location. For example, there are many ways to provide attentional support (checklist items, automated sensors, lights, or other electronic notifications) but not every option can be implemented in every task environment. Further attributes of the user–task–environment system include individual differences in the operator's knowledge, skills, and abilities and the criticality of the task (e.g., Marshall, 2017).

Limitations

There are limitations with using cognitive processes to describe cognitive aids. First, task analysis requires knowledge of cognition beyond many of those currently creating cognitive aids. Thus, translation by experts will be needed once

the taxonomy offers insight into the various ways a task can be supported by cognitive aids. This has been the historical role of human factors professionals, to bridge the knowledge gap of operators or users and designers. A second limitation will be the overlap as to which portions of an aid offer which support. For example, the Surgical Safety Checklist requires the nurse to verbally confirm "completion of instrument, sponge and needle count" before the patient leaves the operating room. This could be argued to support *prospective memory* for the surgeons' intent to remove those items or *recognition memory*, that all items in a tray are indeed in the tray, or *orienting attention*, to ensure that the nurse focused on the tray at the correct time, or all of these combined.

Our last limitation was that only attention aids were fully explained; a similar treatment must be applied to other cognitive processes. Our model provides the structure; research findings and gaps must be established for the other cognitive processes. Research on cognitive aids should include consideration of possible interactions when more than one cognitive process requires support in a task.

Conclusion

A complete understanding of the outcome desired and the demands of the task preceding that outcome is the crucial first step toward developing new cognitive aids. With task analysis, task demands can be matched to the kind of cognitive support required. Using previously classified aids, the types of previously successful support can be nominated for use in a new aid. For example, if the task has prospective memory demands, the aid should support prospective memory using evidence-based designs: alarms, notifications, checklist steps, and look to other aids that were classified as providing prospective memory support for novel ideas.

KEY POINTS

- The literature on cognitive aids can be organized through analysis of what cognitive processes the aid supports.
- Aids can be improved by altering how the aid supports the task or by improving the usability of the aid.
- Considering the individual cognitive processes and interactions being supported by aids provides those making new aids a connection to well-established theories in cognitive psychology.
- The proposed method for future aid design includes thorough analysis of the task to be aided in terms of resource-limited cognitive processes followed by theory and evidence-based design using an organized literature on cognitive support via aids.

ACKNOWLEDGMENTS

This material is based on the work supported by the National Aeronautics and Space Administration (NASA) under Grant No. NNX16AP91G issued through the Human Research Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NASA.

ORCID iD

Anne Collins McLaughlin https://orcid.org/0000-0002-1744-085X

REFERENCES

Abdi, L., Abdallah, F. B., & Meddeb, A. (2015). In-vehicle augmented reality traffic information system: A new type of communication between driver and vehicle. *Procedia Computer Science*, *73*, 242–249. https://doi.org/10.1016/j.procs.2015.12.024

Aldag, R. J., & Power, D. J. (1986). An empirical assessment of computer-assisted decision analysis. *Decision Sciences*, *17*, 572–588. https://doi.org/10.1111/j.1540-5915.1986.tb00243.x

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). Academic Press.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*, 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Barg-Walkow, L. H., & Rogers, W. A. (2017). Modeling task scheduling in complex healthcare environments: Identifying relevant factors. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 61, pp. 772–775). SAGE Publications. https://doi.org/10.1177/1541931213601677

Bodala, I. P., Li, J., Thakor, N. V., & Al-Nashash, H. (2016). EEG and eye tracking demonstrate vigilance enhancement with challenge integration. *Frontiers in Human Neuroscience*, *10*, 273. https://doi.org/10.3389/fnhum.2016.00273

Bosk, C. L., Dixon-Woods, M., Goeschel, C. A., & Pronovost, P. J. (2009). Reality check for checklists. *The Lancet*, *374*, 444–445. https://doi.org/10.1016/S0140-6736(09)61440-9

Broadbent, D. (1958). *Perception and communication*. Pergamon Press.

Burian, B. K., Clebone, A., Dismukes, K., & Ruskin, K. J. (2018). More than a tick box: Medical checklist development, design, and use. *Anesthesia and Analgesia*, *126*, 223–232. https://doi.org/10.1213/ANE.0000000000002286

Chaparro, A., Keebler, J. R., Lazzara, E. H., & Diamond, A. (2019). Checklists: A review of their origins, benefits, and current uses as a cognitive aid in medicine. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *27*, 21–26. https://doi.org/10.1177/1064804618819181

Chung, P. H., & Byrne, M. D. (2008). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. *International Journal of Human-Computer Studies*, *66*, 217–232. https://doi.org/10.1016/j.ijhcs.2007.09.001

Clebone, A., Burian, B. K., & Tung, A. (2019). The effect of cognitive aid design on the perceived usability of critical event cognitive AIDS. *Acta Anaesthesiologica Scandinavica*, *64*, 378–384. https://doi.org/10.1111/aas.13503

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*, 684–687. https://doi.org/10.1126/science.1128356

Degani, A., & Wiener, E. L. (1993). Cockpit checklists: Concepts, design, and use. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*, 345–359. https://doi.org/10.1177/001872089303500209

DeMuth, D. L., & Achorn, E. H. (1978). The physician's income tax checklist. *Pennsylvania Medicine*, *81*, 94–97.

Field, L. C., McEvoy, M. D., Smalley, J. C., Clark, C. A., McEvoy, M. B., Rieke, H., Nietert, P. J., & Furse, C. M. (2014). Use of an electronic decision support tool improves management of simulated in-hospital cardiac arrest. *Resuscitation*, *85*, 138–142. https://doi.org/10.1016/j.resuscitation.2013.09.013

Fletcher, K. A., & Bedwell, W. L. (2014). Cognitive AIDS: Design suggestions for the medical field. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *3*, 148–152.

Folk, C., & Gibson, B. (Eds.). (2001). *Attraction, distraction and action: Multiple perspectives on attentional capture* (Vol. 133). Elsevier Science.

Gibb, R., Schvaneveldt, R., & Gray, R. (2008). Visual misperception in aviation: Glide path performance in a black hole environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*, 699–711. https://doi.org/10.1518/001872008X288619

Goldhaber-Fiebert, S. N., & Howard, S. K. (2013). Implementing emergency manuals: Can cognitive AIDS help translate best practices for patient care during acute events? *Anesthesia and Analgesia*, *117*, 1149–1161. https://doi.org/10.1213/ANE.0b013e318298867a

Gray, W. D., Neth, H., & Schoelles, M. J. (2006). The functional task environment. In A. F. Kramer, D. A. Wiegman, & A. Kirlik (Eds.), *Attention: From Theory to Practice* (pp. 100–118). Oxford University Press.

Hales, B., Terblanche, M., Fowler, R., & Sibbald, W. (2008). Development of medical checklists for improved quality of patient care. *International Journal for Quality in Health Care*, *20*, 22–30. https://doi.org/10.1093/intqhc/mzm062

Haynes, A. B., Weiser, T. G., Berry, W. R., Lipsitz, S. R., Breizat, A. -H. S., Dellinger, E. P., Herbosa, T., Joseph, S., Kibatala, P. L., Lapitan, M. C. M., Merry, A. F., Moorthy, K., Reznick, R. K., Taylor, B., & Gawande, A. A. (2009). A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine*, *360*, 491–499. https://doi.org/10.1056/NEJMsa0810119

Healy, A. F., & Bourne, L. E. (2013). Empirically valid principles for training in the real world. *The American Journal of Psychology*, *126*, 389–399. https://doi.org/10.5406/amerjpsyc.126.4.0389

Helin, K., Kuula, T., Vizzi, C., Karjalainen, J., & Vovk, A. (2018). User experience of augmented reality system for astronaut's manual work support. *Frontiers in Robotics and AI*, *5*, 1–10. https://doi.org/10.3389/frobt.2018.00106

Hepner, D. L., Arriaga, A. F., Cooper, J. B., Goldhaber-Fiebert, S. N., Gaba, D. M., Berry, W. R., Boorman, D. J., & Bader, A. M. (2017). Operating room crisis checklists and emergency manuals. *Anesthesiology*, *127*, 384–392. https://doi.org/10.1097/ALN.0000000000001731

Hormann, A. M. (1971). A man-machine synergistic approach to planning and creative problem solving. Part 1. *International Journal of Man-Machine Studies*, *3*, 167–184. https://doi.org/10.1016/S0020-7373(71)80013-5

Howard, S. K., Chu, L. K., Goldhaber-Fiebert, S. N., Gaba, D. M., & Harrison, T. K. (2013). *Emergency manual: Cognitive aids for perioperative clinical events*. Stanford Anesthesia Cognitive Aid Group, Creative Commons BY-NC-ND. http://emergencymanual.stanford.edu

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kapur, N., Parand, A., Soukup, T., Reader, T., & Sevdalis, N. (2015). Aviation and healthcare: A comparative review with implications for patient safety. *JRSM Open*, *7*, 1–10. https://doi.org/10.1177/2054270415616548

Kim, S., & Dey, A. K. (2009). Simulated augmented reality windshield display as a cognitive mapping aid for elder driver navigation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 133–142). https://doi.org/10.1145/1518701.1518724

Kobayashi, L., Gosbee, J. W., & Merck, D. L. (2017). Development and application of a clinical microsystem simulation methodology for human factors-based research of alarm fatigue. *HERD: Health Environments Research & Design Journal*, *10*, 91–104. https://doi.org/10.1177/1937586716673829

Long, E., Cincotta, D., Grindlay, J., Pellicano, A., Clifford, M., & Sabato, S. (2017). Implementation of NAP4 emergency airway management recommendations in a quaternary-level pediatric hospital. *Pediatric Anesthesia*, *27*, 451–460. https://doi.org/10.1111/pan.13128

Marmaras, N., & Kontogiannis, T. (2001). Cognitive tasks. In G. Salvendy (Ed.), *Handbook of industrial engineering* (pp. 1011–1040). John Wiley and Sons Ltd.

Marshall, S. D. (2013). Use of cognitive AIDS during emergencies in anesthesia: A systematic review. *Anesthesia and Analgesia*, *117*, 1162–1171.

Marshall, S. D. (2017). Helping experts and expert teams perform under duress: An agenda for cognitive aid research. *Anaesthesia*, *72*, 289–295. https://doi.org/10.1111/anae.13707

McLaughlin, A. C., & Byrne, V. (2019). *Human factors in cognitive aid design and use* (Technical Report NNX16AP91G). North Carolina State University, Department of Psychology.

McLaughlin, A. C., Ward, J., & Keene, B. W. (2016). Development of a veterinary surgical checklist. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *24*, 27–34. https://doi.org/10.1177/1064804615621411

Miller, S. L., Adelman, L., de Henderson, E. V., Schoelles, M., & Yeo, C. (2000). Team decision-making strategies: Implications for designing the interface in complex tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*, 81–84. https://doi.org/10.1177/154193120004400122

Minotra, D., & McNeese, M. D. (2017). Predictive AIDS can lead to sustained attention decrements in the detection of non-routine critical events in event monitoring. *Cognition, Technology & Work*, *19*, 161–177. https://doi.org/10.1007/s10111-017-0402-x

Müller, H. J., & Rabbitt, P. M. (1989). Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 315–330. https://doi.org/10.1037/0096-1523.15.2.315

Myers, P. (2016). Commercial aircraft electronic checklists: Benefits and challenges (literature review). *International Journal of Aviation, Aeronautics, and Aerospace*, *3*, 1–10. https://doi.org/10.15394/ijaaa.2016.1112

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, *86*, 214–255. https://doi.org/10.1037/0033-295X.86.3.214

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books.

Rall, M., Gaba, D., Howard, S., & Dieckmann, P. (2010). Human performance and patient safety. In R. Miller, L. Eriksson, & L. Fleisher (Eds.), *Miller's anesthesia* (8th ed.). Elsevier.

Ralph, B. C. W., Thomson, D. R., Cheyne, J. A., & Smilek, D. (2014). Media multitasking and failures of attention in everyday life. *Psychological Research*, *78*, 661–669. https://doi.org/10.1007/s00426-013-0523-7

Ramachandran, M., Greenstein, J. S., McEvoy, M., & McEvoy, M. D. (2014). Using a context-sensitive ranking method to organize reversible causes of cardiac arrest in a digital cognitive aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*, 788–792. https://doi.org/10.1177/1541931214581144

Ratwani, R. M., & Trafton, J. G. (2011). A real-time eye tracking system for predicting and preventing postcompletion errors. *Human–Computer Interaction*, *26*, 205–245.

Reason, J. (1987). Cognitive AIDS in process environments: Prostheses or tools? *International Journal of Man-Machine Studies*, *27*, 463–470. https://doi.org/10.1016/S0020-7373(87)80010-X

Renna, T. D., Crooks, S., Pigford, A.-A., Clarkin, C., Fraser, A. B., Bunting, A. C., Bould, M. D., & Boet, S. (2016). Cognitive AIDS for role definition (CARD) to improve interprofessional team crisis resource management: An exploratory study. *Journal of Interprofessional Care*, *30*, 582–590. https://doi.org/10.1080/13561820.2016.1179271

Rill, R. A., Faragó, K. B., & Lőrincz, A. (2018). Strategic predictors of performance in a divided attention task. *PLoS One*, *13*, 1–27.

Rusch, M. L., Schall, M. C., Gavin, P., Lee, J. D., Dawson, J. D., Vecera, S., & Rizzo, M. (2013). Directing driver attention with augmented reality cues. *Transportation Research Part F: Traffic Psychology and Behaviour*, *16*, 127–137. https://doi.org/10.1016/j.trf.2012.08.007

Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., & Wiegmann, D. A. (2007). Human error and commercial aviation accidents: An analysis using the human factors analysis and classification system. *Human Factors*, *49*, 227–242. https://doi.org/10.1518/001872007X312469

Stern, E. (2017). Individual differences in the learning potential of human beings. *npj Science of Learning*, *2*, 2. https://doi.org/10.1038/s41539-016-0003-0

Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention. *The American Journal of Psychology*, *77*, 206–219. https://doi.org/10.2307/1420127

Weiss, M. J., Bhanji, F., Fontela, P. S., & Razack, S. I. (2013). A preliminary study of the impact of a handover cognitive aid on clinical Reasoning and information transfer. *Medical Education*, *47*, 832–841. https://doi.org/10.1111/medu.12212

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*, 159–177. https://doi.org/10.1080/14639220210123806

Wickens, C. D., & Gutzwiller, R. S. (2017). The status of the strategic task overload model (STOM) for predicting multitask management. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 61, pp. 757–761). SAGE Publications. https://doi.org/10.1177/1541931213601674

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology and human performance*. Psychology Press.

Wiener, E. L. (1977). Controlled flight into terrain accidents: System-induced errors. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *19*, 171–181. https://doi.org/10.1177/001872087701900207

Winters, B. D., Gurses, A. P., Lehmann, H., Sexton, J. B., Rampersad, C. J., & Pronovost, P. J. (2009). Clinical review: Checklists - translating evidence into practice. *Critical Care*, *13*, 210. https://doi.org/10.1186/cc7792

World Health Organization (WHO). (2015). *Team already knows each other*. Safe Surgery 2015. (August 22, 2019). http://www.safesurgery2015.org/uploads/1/0/9/0/1090835/safe_surgery_checklist_template_-_team_knows_each_other__rev_07aug15_.pdf

Wu, L., Cirimele, J., Leach, K., Card, S., Chu, L., Harrison, T. K., & Klemmer, S. R. (2014). Supporting crisis response with dynamic procedure aids. In *Proceedings of the 2014 conference on designing interactive systems* (pp. 315–324). ACM.

Anne Collins McLaughlin is currently a professor in the Department of Psychology at North Carolina State University in Raleigh, NC. She earned her PhD in psychology in 2007 from the Georgia Institute of Technology. Her research interests include the study of individual differences in cognition, particularly those that tend to change with age, applied to various domains including training and cognition aids.

Vicky E. Byrne is currently a senior Human Factors Engineer at KBR in Houston, TX. She earned her MS in psychology in 1993 from the Georgia Institute of Technology. She has worked on a multitude of projects in the space domain regarding the human factors of interface design, instructional design, training, and healthcare.

Date received: September 11, 2019
Date accepted: March 25, 2020

The Benefits and Costs of Low and High Degree of Automation

Monica Tatasciore, **Vanessa K. Bowden**, **Troy A. W. Visser**, **Steph I. C. Michailovs**, and **Shayne Loft**, The University of Western Australia, Perth, Australia

Objective: The objective of this study is to examine the effects of low and high degree of automation (DOA) on performance, subjective workload, situation awareness (SA), and return-to-manual control in simulated submarine track management.

Background: Theory and meta-analytic evidence suggest that as DOA increases, operator performance improves and workload decreases, but SA and return-to-manual control declines. Research also suggests that operators have particular difficulty regaining manual control if automation provides incorrect advice.

Method: Undergraduate student participants completed a submarine track management task that required them to track the position and behavior of contacts. Low DOA supported information acquisition and analysis, whereas high DOA recommended decisions. At a late stage in the task, automation was either unexpectedly removed or provided incorrect advice.

Results: Relative to no automation, low DOA moderately benefited performance but impaired SA and non-automated task performance. Relative to no automation and low DOA, high DOA benefited performance and lowered workload. High DOA did impair non-automated task performance compared with no automation, but this was equivalent to low DOA. Participants were able to return-to-manual control when they knew low or high DOA was disengaged, or when high DOA provided incorrect advice.

Conclusion: High DOA improved performance and lowered workload, at no additional cost to SA or return-to-manual performance when compared with low DOA.

Application: Designers should consider the likely level of uncertainty in the environment and the consequences of return-to-manual deficits before implementing low or high DOA.

Keywords: automation, submarine track management, situation awareness, workload, complacency

Address correspondence to Monica Tatasciore, The University of Western Australia, 35 Stirling Highway, Perth, Western Australia 6009, Australia; e-mail: monica.tatasciore@research.uwa.edu.au.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 874–896
DOI: 10.1177/0018720819867181
Article reuse guidelines: sagepub.com/journals-permissions

Technological developments in computer hardware and software have made it possible to automate many aspects of complex work systems, significantly improving workplace efficiency and safety (Sheridan, 2015; Vagia, Transeth, & Fjerdingen, 2016). Automation can be defined as "a device or system that accomplishes a function that was previously, or conceivably could be, carried out by a human operator" (Parasuraman, Sheridan, & Wickens, 2000, p. 287). Examples of automation include image-guided navigation tools in surgery, flight management systems in cockpits, aircraft separation assurance technology in air traffic control, and decision aids in unmanned vehicle control.

Whereas routine lower level tasks have typically been the first to be automated due to their operational predictability, with ongoing emphasis on maximizing system capacity, and the development of sophisticated machine-learning algorithms, automation can now begin to recommend or even execute high-level decisions for operators. The submarine control room is one area in which this type of "decision-level" automation is rapidly developing (Roberts, Stanton, & Fay, 2017). Submarine track management, for example, requires operators to coordinate information across multiple displays to create a tactical picture of the position and behavior of contacts in relation to the submarine (Ownship) and strategic landmarks (Kirschenbaum, 2011). A key question in this, and in similar work contexts (e.g., unmanned vehicle control, air traffic control), concerns the extent to which an operator can effectively use automated systems that recommend decisions.

Researchers have long recognized the potential costs associated with automation. These costs include reductions in operators' understanding of

a task and their ability to anticipate future task events (situation awareness [SA]; Endsley, 1988) due to automation-induced complacency (Parasuraman, Molloy, & Singh, 1993; Parasuraman & Riley, 1997), and reductions in operators' ability to regain manual control after automation use (Kaber & Endsley, 2004; Parasuraman & Manzey, 2010). Notably, there is also some evidence that these costs increase as automation begins to assume higher level functions. For example, Onnasch, Wickens, Li, and Manzey (2014) reviewed 18 studies that varied in "degree of automation" (DOA)—an ordinal metric that ranked the level of work the automation was doing (Sheridan & Verplank, 1978) across four stages: information acquisition, information analysis, decision recommendation, and action execution (Parasuraman et al., 2000). Onnasch et al. (2014) found that as DOA increased, performance improved and workload decreased. However, SA and return-to-manual performance declined (for examples of specific studies that have shown this trade-off, see Kaber, Onal, & Endsley, 2000; Li, Wickens, Sarter, & Sebok, 2014; Manzey, Reichenbach, & Onnasch, 2012).

More recently, Chen, Visser, Huf, and Loft (2017) asked participants to monitor a submarine track management tactical display ("Surface Plot") that presented the location and heading of contacts in relation to the Ownship and landmarks, and a "waterfall" display that presented sonar bearings of contacts and how those bearings changed with time. Participants performed three tasks. The classification task required participants to classify contacts (hostile, friendly, etc.) based on how long they had spent within certain display regions. The closest point of approach (CPA) task required participants to monitor changes in contact heading to determine their CPA to Ownship. The dive task required participants to integrate contact location and heading information to determine when the submarine could safely dive. The simulation automated information acquisition and analysis stages of the classification and CPA tasks (i.e., relatively low DOA) by indicating to participants when contacts first entered display regions (to aid contact classification) and tracked when contacts made heading changes (to aid CPA detection).

Chen et al. (2017; Experiment 3; between-subjects design) demonstrated that low DOA resulted in benefits to classification performance (accuracy and response time [RT]) but not CPA performance, and did not reduce subjective workload, compared with when no automation was provided. In addition, participant SA was poorer when automation was used, as was performance on the non-automated dive task. The cost observed to the non-automated dive task with the use of automation is critical to further explore because this novel finding suggests that operators in complex work systems may find it difficult to maintain adequate performance on non-automated tasks that share information processing requirements with currently automated tasks. After low DOA was unexpectedly removed, costs to SA did not diminish, although there were no associated return-to-manual performance deficits.

With defense and other industries focused on developing high DOA that recommends decisions to operators (Endsley, 2017; U.S. Air Force, 2015), it is critical to further understand how high DOA systems can affect operators. Under conditions when automation is reliable, high DOA that recommends decisions to operators should further improve performance and reduce workload compared with low DOA (Onnasch et al., 2014). The key question concerns whether high DOA comes at increased cost to concurrent non-automated task performance and SA, or return-to-manual performance, compared with low DOA. The answer to this is critical for work design. If high DOA produces greater benefit at no extra cost, then it would be more desirable to employ than low DOA. If, however, high DOA produces greater benefit but at extra cost, whether high DOA is deployed would depend on factors such as the level of uncertainty in the environment or the operational consequences of reduced concurrent non-automated task performance, loss of SA, or return-to-manual performance deficits (Endsley, 2017; Wickens, Clegg, Vieane, & Sebok, 2015).

With these questions in mind, the current study began by examining the effects of low DOA and high DOA on operator performance, workload, SA, non-automated task performance,

and return-to-manual performance in submarine track management. Low DOA was identical to that used by Chen et al. (2017) and supported information acquisition and analysis by displaying when and for how long contacts were positioned in an area of interest (classification task) and by displaying contact heading changes (CPA task). High DOA not only provided the same information acquisition and analysis information but also made explicit recommendations to participants regarding when and what to classify contacts, and when a contact had made a CPA. The purpose of Experiment 1 was (a) to replicate the effects of low DOA on performance and SA that were demonstrated by Chen et al. (2017) when compared with no automation and (b) to examine whether high DOA produces benefits to performance and workload compared with no automation and low DOA, and whether high DOA increases costs to non-automated task performance, SA, or return-to-manual performance compared with no automation and low DOA. Our predictions regarding the effects of DOA are summarized in Table 1 and described in detail later.

AUTOMATED TASK PERFORMANCE

The classification and CPA tasks were automated for the low and high DOA conditions. Performance on these two tasks was assessed by accuracy and RT. Under routine states (when automation was reliably functioning), we expected higher classification accuracy and faster classification RT with increasing DOA. While Chen et al. (2017) did not show a benefit to CPA accuracy with the use of low DOA under routine states, we expected to find benefits to CPA accuracy with high DOA. Furthermore, Chen et al. found that participants made slower CPA decisions when using low DOA, which reflects that the automated track history allowed participants to detect CPAs well after they had occurred by detecting past heading changes. In contrast, high DOA should allow participants to make faster CPA decisions compared with both no automation and low DOA because it highlights CPA events at the actual time they occur.

Chen et al. (2017) found no return-to-manual deficits to the classification or CPA tasks when low DOA was removed. However, the Onnasch et al. (2014) meta-analysis indicated that the negative consequences of automation are more likely with higher DOA. From a theoretical perspective, higher DOA that recommends decisions could reduce the perceived need to actively process raw information (e.g., contact position and heading) on the displays (complacency; Parasuraman & Manzey, 2010; Wickens, Sebok, Li, Sarter, & Gacy, 2015). Theory and evidence from the broader psychological science literature also predicts poorer understanding and retention of information when individuals passively process information rather than actively making decisions (e.g., the generation effect, Slamecka & Graf, 1978; the testing effect, Roediger & Karpicke, 2006; transfer of training, Blume, Ford, Baldwin, & Huang, 2010). We therefore expected to find deficits to performance on the classification and CPA tasks when high DOA was removed, as compared with the no automation and low DOA conditions.

NON-AUTOMATED TASK PERFORMANCE

The Chen et al. (2017) cost observed to the non-automated dive task (to both accuracy and RT) under routine states with the use of low DOA suggests that participants found it difficult to maintain performance on the non-automated task that shared information processing requirements with the automated tasks. That is, dive task performance was degraded because participants scrutinized contact location and heading information less closely when using automation (complacency). To the extent that complacency effects are heightened with high DOA as suggested by Onnasch et al. (2014), we would expect dive task performance to be further impaired with the use of high DOA. Furthermore, on the basis of Chen et al.'s (2017) and Onnasch et al.'s (2014) meta-analytic evidence, we expected return-to-manual deficits when high DOA was removed, as compared with the no automation and low DOA conditions.

WORKLOAD

Chen et al. (2017) did not find reduced subjective workload with low DOA during routine

TABLE 1: Predictions for Experiment 1

Task		Routine	Removal
Classification	Accuracy	None < Low < High (the higher the DOA, the better the accuracy)	[None = Low] > High (lower accuracy after high DOA removal)
	RT	None > Low > High (the higher the DOA, the faster the decisions)	[None = Low] < High (slower decisions after high DOA removal)
CPA	Accuracy	[None = Low] < High (benefits to accuracy with high DOA)	[None = Low] > High (lower accuracy after high DOA removal)
	RT	High < None < Low (benefits to RT with high DOA, slower decisions with low DOA)	[None = Low] < High (slower decisions after high DOA removal)
Dive	Accuracy	None > Low > High (the higher the DOA, the poorer the accuracy)	[None = Low] > High (lower accuracy after high DOA removal)
	RT	None < Low < High (the higher the DOA, the slower the decisions)	[None = Low] < High (slower decisions after high DOA removal)
Workload		[None = Low] > High (reduced workload with high DOA)	[None = Low] < High (higher workload after high DOA removal)
SA		None > Low > High (the higher the DOA, the poorer the SA)	None > Low > High (after removal, the higher the DOA, the poorer the SA)

Note. Routine = automation is reliable; Removal = after automation is removed; DOA = degree of automation; RT = response time; None = no automation; Low = low DOA; High = high DOA; CPA = closest point of approach; SA = situation awareness.

states, but on the basis of the Onnasch et al. (2014) meta-analysis, we expected reduced subjective workload with high DOA, as compared with the no automation and low DOA conditions. Chen et al. (2017) found no return-to-manual increase in workload when low DOA was removed, but on the basis of Onnasch et al. (2014), we expected to find increased subjective workload when high DOA was removed, as compared with the no automation and low DOA conditions.

SA

Chen et al. (2017) found reduced SA with low DOA during routine states, and based on Onnasch et al. (2014), we expected SA to be further impaired with high DOA. Chen et al. (2017) found reduced SA when low DOA was removed, and on the basis of Onnasch et al. (2014), we expected to find that SA would be further impaired when high DOA was removed.

EXPERIMENT 1

Participants

Participants were 122 (86 females) undergraduate psychology students (age: M = 23 years, SD = 7.2) who took part for course credit and were randomly assigned to one of three conditions: no automation (n = 42), low DOA (n = 40), and high DOA (n = 40). This research complied with the American Psychological Association Code of Ethics and was approved by the Human Research Ethics Office at the University of Western Australia. Informed consent was obtained from each participant.

Design

A mixed design was used, where the between-subjects factor was condition (no automation, low DOA, high DOA) and the within-subjects factor was automation state (routine, automation removal). Automation condition was manipulated between-subjects so that each participant only experienced the unexpected automation failure once to ensure that there were no carryover effects (first-failure effect; see Merlo, Wickens, & Yeh, 2000). Participants completed three 27.5-min track management scenarios, each corresponding to different Australian costal maps.

Simulated Submarine Track Management Task

The track management simulation (Figure 1) was developed based on a task analysis conducted with Royal Australian Navy Submariners (Chen et al., 2017). The tactical display, presented on the left monitor, showed a "bird's eye" view of the area with concentric rings representing distance from the center point (Ownship). This tactical display presented the location and heading of contacts. The waterfall display, presented on the right monitor, showed contact bearings in relation to Ownship on the top horizontal axis, and how these bearings changed with time along the vertical axes. This information was displayed as vertical lines or "soundtracks," which grew downward with time. Task load periodically varied with the number of contacts increasing (maximum of eight contacts) and decreasing (minimum of one contact) 3 times during each 27.5-min scenario.

The "Track Assist" automation interface was located at the bottom right of the tactical display and allowed participants to determine whether the automation was always on (fixed) or not available (none). During the third scenario, automation was unexpectedly removed (10.58-min into the 27.5-min scenario). When the automation was removed, a message appeared on the tactical display: "Attention. ENEMY SONAR detected. Track Assist turned off. Manual tracking required." Participants were required to acknowledge this message by clicking an "ok" button. In the no automation condition, a message was presented at the same time that read: "Attention. ENEMY SONAR detected. Keep vigilant and continue to track vessels."

Classification task. Participants classified contacts depending on how long they spent within specific areas on the tactical display. A contact was a "Friendly" if it spent more than 2 continuous minutes within the area bounded by blue lines on the tactical display. A contact was a "Merchant" if it spent more than 2 continuous minutes within the "shipping lane" represented by two white parallel lines on the tactical display. A contact was a "Trawler" if it spent more than 2 continuous minutes in the shallow dark blue areas on the tactical display. A contact was an "Enemy" if in the first 4-min of its presentation, it had not spent at least 1 continuous minute in any classification zone. To track whether a contact had been in a given area for more than 2 min, participants could place horizontal lines on the top of each soundtrack on the waterfall display when a contact entered an area of interest. When this line reached the 2-min mark, a contact could be classified. To detect enemies, participants could place the horizontal line on the bottom of the soundtrack of any contact that had not crossed into an area of interest. Once this horizontal line reached 4-min, the contact could be classified as an enemy.

The contact classification task could be automated to either a low or high degree. For low DOA, horizontal lines were automatically placed on the soundtrack when a contact entered an area of interest. In addition, a horizontal line was automatically placed at the bottom of the soundtrack when it reached the 4-min mark to assist with classifying enemies. Participants still had to monitor the horizontal lines to see when they reached the 2-min mark (or 4-min mark for enemies) to classify contacts. When automation was removed in the third scenario, the existing horizontal lines on the waterfall display remained, but subsequent lines had to be manually entered.

High DOA was identical to low DOA, except that a square box with the recommended classification (i.e., f = Friendly, m = Merchant, t = Trawler, e = Enemy; see Figure 1) was attached to the horizontal lines on the soundtracks. In addition, when the horizontal line reached the 2-min mark, it flashed to notify participants that the contact had been in an area of interest for 2

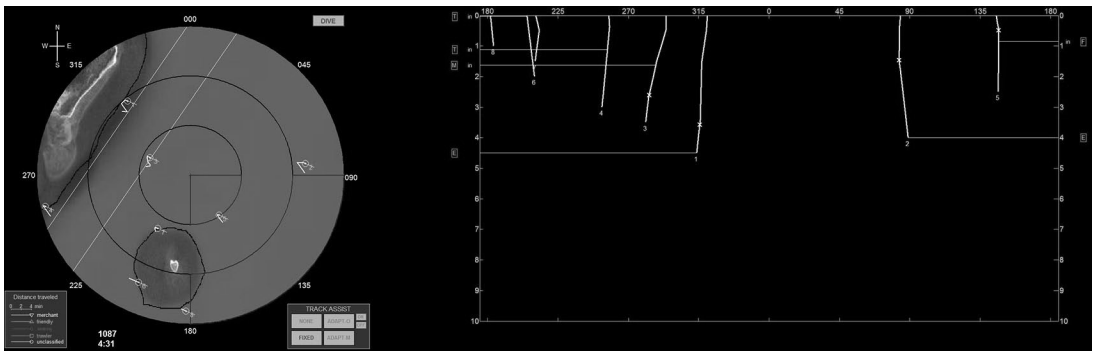


Figure 1. An example of submarine track management scenario. The display on the left is the tactical display which presents a bird's eye view of the area with concentric rings representing distance from the Ownship. The display on the right is the waterfall display, which provides the bearing of contacts in relation to the Ownship on the top horizontal axis, and how those bearings change with time along the vertical axes. These data are presented as soundtracks that grow downward with time. Eight contacts are displayed. Projecting from the center of each contact is a line which indicates the current heading of the contact. In this example, high DOA is active. On the tactical display, a track history is attached to each contact to reflect contact heading changes to assist with the CPA task. The track history will flash to notify participants when a CPA has occurred. Present on the waterfall display are horizontal lines which are automatically placed when a contact enters an area of interest. Attached to these lines are boxes which include the appropriate classification letter of a given contact. These lines will flash when a contact can be classified. The low DOA displays are the same to that shown in this figure, except there will be no boxes with the appropriate classification letter attached to the horizontal lines. In addition, the track history and the horizontal lines will not flash to signal a classification or CPA event with the use of low DOA. When no automation is provided, contacts will not have a track history for the CPA task, and horizontal lines must be placed manually for the classification task.

min. If the contact was an enemy, the horizontal line flashed at the 4-min mark. It remained the participants' task to then execute the classification task action (or not) after receiving the automated advice. When automation was removed, the existing horizontal lines and any classification letters remained, but did not flash in the future, and subsequent lines had to be manually entered.

CPA. The CPA is defined as the time at which a contact that was heading toward the Ownship turned away from the Ownship. Participants reported the time at which the CPA occurred by placing a cross on the corresponding soundtrack on the waterfall display. Each contact had one CPA per scenario. For the two automation conditions, the CPA task was automated to a low or high degree. For low DOA, each contact was presented with a track history, which reduced the need for participants to track which contacts made heading changes. Participants were still required to interpret the track history to mark the timing of each CPA on the waterfall display. For high DOA, the track history also flashed to alert the participant when the contact had turned away from the Ownship. It was then the participant's responsibility to mark the appropriate CPA time on the waterfall display. When automation was removed, the existing track history for high DOA and low DOA remained on the tactical display but was not updated to reflect further contact movement.

Dive task. Participants were required to dive when (a) all contacts on the tactical display were heading in the same direction and (b) one contact was heading directly toward the Ownship. There were either 9 or 10 dive windows per scenario, and each dive window varied in duration between 10 and 30 s. Participants responded to dive windows by pressing the dive button. The dive task was not automated.

TABLE 2: The SAGAT Queries Used to Measure Participant SA

SA Level	SAGAT Queries		
1	Which vessel is currently in an X zone?	How many vessels are heading away from you?	How many vessels are currently facing the same direction?
	Is vessel X currently in an X zone?	Is vessel X heading away from you?	Are any vessels heading directly toward you?
			How many vessels are heading away from you?
2	Has any vessel been in an X zone for more than 1 min?	How many times has vessel X changed course?	Are any vessels heading in the same direction?
	How many vessels are currently in an X zone?	Has vessel X had any kinks in its soundtrack?	Which vessel is currently heading toward you?
	Which vessel most recently crossed a classification boundary?		
3	Which unclassified vessel is most likely to be an X?	Which vessel would make a CPA if it turned to a heading of xxx?	Would vessel X head directly toward you if it turned to a heading of xxx?
	Could vessel X cross a boundary within 4-min time?	Would a CPA be made for vessel X if it turned to a heading of xxx?	

Note. SAGAT = Situation Awareness Global Assessment Technique; SA = situation awareness; CPA = closest point of approach.

Measures

SA. SA was measured using the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995). During each scenario, the simulation was paused 6 times, and the tactical and waterfall displays were blanked and replaced with SAGAT queries. Within each freeze, seven SAGAT queries were delivered. The first question always asked participants to mark a specific contact location on the tactical display, whereas the remaining six questions targeted knowledge necessary for the classification, CPA, and dive tasks, at the three levels of SA (Endsley, 1995). During each SAGAT freeze, all participants received the same six SA queries which were taken from the pool of queries presented in Table 2.

Workload. Two subjective measures of workload were used. The Air Traffic Workload Input Technique (ATWIT; Stein, 1985) was presented on the tactical display every minute throughout the scenario. Participants had 10 s to select a workload rating between 1 and 10 (1–2 = very low, 3–5 = moderate, 6–8 = relatively high, and 9–10 = very high). The National Aeronautics and Space Administration Task Load Index (NASA-TLX; Hart & Staveland, 1987) was completed after each scenario. A NASA-TLX score was calculated by multiplying the ratings for each subscale of workload by its corresponding weighting, adding the values for all the subscales together, and then dividing the total by 15.

Procedure

The experiment duration was 3 hr. First, participants completed an 80-min training session. Training began with a 35-min audiovisual PowerPoint presentation that explained the task and included "learning checks." Following this, participants viewed a narrated video of the simulation in which all tasks were demonstrated

without automation. Participants then completed a 27.5-min practice scenario with no automation. Participants who were in either the low or high DOA conditions watched a PowerPoint presentation that explained their automation. Participants then completed three 27.5-min experimental scenarios in their assigned automation condition. Each scenario contained unique contacts and the order of scenario maps was counterbalanced.

Results

The hit rates for the classification, CPA, and dive tasks were calculated as the number of correct task responses per scenario divided by the total number of task events. RTs were based on correct decisions only. A CPA was marked as correct if the cross was placed at any time 1.5 s before or after the actual CPA, so long as the cross was placed on the correct soundtrack. If placed outside this temporal range, then the cross was recorded as a false alarm. A parameter was estimated for false alarm rates as the exact number of contacts and events associated with making a false alarm was indeterminable (Chen et al., 2017). CPA false alarms were most likely to be made in response to contact course changes. The false alarm rate was therefore calculated to be the number of false alarms divided by the number of course changes, minus 24 (the total number of CPA events per scenario). CPA performance was then calculated by subtracting the CPA false alarm rate from the hit rate. Course changes were always required for a dive window to be open; hence, a dive false alarm was most likely to be made during a course change. As there are fewer dive windows than CPAs, as well as the rule that all contacts need to be heading in the same direction for a dive window, it was less likely that every course change could be mistaken for a dive window. Therefore, the dive false alarm rate was calculated by dividing the number of false alarms by half the number of course changes, minus the total number of dive windows (Chen et al., 2017). Dive task performance was then calculated by subtracting the dive false alarm rate from the hit rate.

The means and between-subjects 95% confidence intervals (CIs) for performance, subjective workload, and SA are presented in Table 3. Data are separated into the time period automation was available (routine state: first two scenarios and first third of the last scenario) and the time period automation was removed (automation removal state: last two thirds of the last scenario). To test our predictions, we ran 3 Condition (no automation, low DOA, high DOA) × 2 Automation State (routine, automation removal) mixed analyses of variance (ANOVAs) on the performance, workload, and SA variables. The between-subjects factor was condition and the within-subjects factor was automation state. The ANOVA results are summarized in Table 4. Significant main effects of condition, or interactions between condition and automation state, were followed up with tests of simple effects (reported in text). To do this, we ran one-way ANOVAs separately for the routine and automation removal states. We then followed significant one-way ANOVAs with post hoc *t* tests that compared the three conditions (no automation, low DOA, high DOA) to each other, correcting for family-wise error by reporting Bonferroni-corrected *p* values (the actual *p* value was multiplied by the number of comparisons for each dependent variable, which was three). Estimates of Cohen's *d* suggested we had a power of 0.82 to detect the medium-to-large effect sizes previously reported by Chen et al. (2017; Cohen, 1988).

Note that several main effects of automation state were found (see Table 4). Classification and CPA performance was poorer, and workload higher, after automation removal compared with routine states. Similar effects were reported by Chen et al. (2017) and are likely caused by the fact that for the low and high DOA conditions (which constitute two of the three conditions and thus 2/3 of the data), the removal state represented the time that automation was removed. For brevity, the main effects of automation state are not further discussed, and we focus on following up the main effect of condition and the interactions.

Automated Task Performance

Classification task. For classification accuracy, there was a main effect of condition and a Condition × State interaction. The simple effect test revealed a significant difference between the

TABLE 3: Descriptive Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 1

	Classification		CPA		Dive		SAGAT	Workload Rating	
Automation	Hit	RT	Hit-FA	RT	Hit-FA	RT	Accuracy	ATWIT	NASA-TLX
Routine state									
None	0.70 [0.62, 0.78]	29.40 [26.01, 32.91]	0.25 [0.18, 0.32]	18.70 [13.76, 23.64]	0.70 [0.64, 0.76]	9.08 [8.06, 10.09]	0.58 [0.54, 0.61]	4.83 [4.47, 5.18]	59.30 [54.87, 63.73]
Low	0.77 [0.70, 0.85]	23.60 [20.64, 26.56]	0.30 [0.21, 0.40]	32.62 [25.53, 39.72]	0.55 [0.48, 0.63]	10.19 [8.98, 11.40]	0.50 [0.45, 0.54]	4.83 [4.44, 5.23]	61.35 [56.70, 66.00]
High	0.90 [0.85, 0.95]	19.32 [16.30, 22.35]	0.74 [0.64, 0.84]	12.31 [10.07, 15.01]	0.57 [0.50, 0.65]	10.90 [9.84, 11.96]	0.52 [0.48, 0.56]	4.10 [3.70, 4.50]	54.02 [48.72, 59.33]
Automation removal state									
None	0.68 [0.59, 0.78]	26.10 [20.09, 32.11]	0.30 [0.22, 0.38]	18.69 [12.65, 24.74]	0.73 [0.66, 0.80]	8.71 [6.99, 10.43]	0.51 [0.47, 0.56]	4.86 [4.41, 5.32]	57.68 [51.63, 63.74]
Low	0.65 [0.56, 0.75]	31.12 [24.70, 37.54]	0.25 [0.17, 0.33]	30.89 [12.57. 49.21]	0.72 [0.64, 0.81]	9.77 [8.51, 11.03]	0.53 [0.49, 0.56]	5.92 [5.44, 6.41]	66.99 [61.44, 72.54]
High	0.79 [0.72, 0.86]	23.36 [19.58, 27.14]	0.44 [0.36, 0.52]	18.37 [12.70, 24.05]	0.71 [0.62, 0.81]	9.86 [8.11, 11.61]	0.54 [0.49, 0.59]	5.24 [4.71, 5.77]	60.63 [55.13, 66.14]

Note. The 95% between-subjects confidence intervals are presented in brackets. CPA = closest point of approach; SAGAT = Situation Awareness Global Assessment Technique; RT = response time; FA = false alarm; ATWIT = Air Traffic Workload Input Technique; NASA-TLX = National Aeronautics and Space Administration Task Load Index.

TABLE 4: Inferential Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 1

Dependent Variable	Effect	*F*	*df*	*p*	η_p^2
Classification (Hit)	Condition	5.38	(1, 119)	.01*	.08
	State	26.69	(1, 119)	<.001*	.18
	Interaction	4.49	(1, 119)	.01*	.07
Classification (RT)	Condition	3.89	(1, 115)	.02*	.06
	State	3.46	(1, 115)	.07	.03
	Interaction	4.50	(1, 115)	.01*	.07
CPA (Hit-FA)	Condition	21.95	(1, 80)	<.001*	.27
	State	29.84	(1, 119)	<.001*	.20
	Interaction	31.25	(1, 119)	<.001*	.34
CPA (RT)	Condition	5.83	(1, 113)	.01*	.09
	State	0.56	(1, 113)	.46	.01
	Interaction	0.50	(1, 113)	.61	.01
Dive (Hit-FA)	Condition	1.53	(1, 119)	.22	.03
	State	40.63	(1, 119)	<.001*	.26
	Interaction	6.02	(1, 119)	.003*	.09
Dive (RT)	Condition	2.21	(1, 117)	.14	.04
	State	1.85	(1, 117)	.18	.02
	Interaction	0.19	(1, 117)	.83	.003
NASA-TLX	Condition	2.19	(1, 118)	.12	.04
	State	13.01	(1, 118)	<.001*	.10
	Interaction	7.03	(1, 118)	.001*	.11
ATWIT	Condition	3.31	(1, 117)	.04*	.05
	State	62.43	(1, 117)	<.001*	.35
	Interaction	14.39	(1, 117)	<.001*	.20
SAGAT (Accuracy)	Condition	1.04	(1, 117)	.36	.02
	State	0.14	(1, 117)	.71	.001
	Interaction	7.29	(1, 117)	.001*	.11

Note. RT = response time; CPA = closest point of approach; FA = false alarm; NASA-TLX = National Aeronautics and Space Administration Task Load Index; ATWIT = Air Traffic Workload Input Technique; SAGAT = Situation Awareness Global Assessment Technique.
*$p < .05$.

conditions during routine states, $F(2, 119) = 9.04$, $p < .001$, $\eta^2 = .13$. During routine states, there was no difference in classification accuracy between the no automation and low DOA conditions, $t < 1$. However, participants provided high DOA made more accurate classifications than participants provided no automation, $t(80) = 4.43$, $p < .001$, $d = 0.99$, or low DOA, $t(78) = 2.84$, $p = .02$, $d = 0.64$. For automation removal states, the simple effect test indicated no significant difference between the conditions, $F(2, 119) = 2.85$, $p = .06$, $\eta^2 = .05$. In summary, only high DOA benefited classification accuracy during routine states, and there were no return-to-manual deficits to classification accuracy following the use of low or high DOA.

For classification RT, there was a main effect of condition and a Condition × State interaction. The simple effect test revealed a significant difference between the conditions during routine

states, $F(2, 118) = 10.66$, $p < .001$, $\eta^2 = .15$. During routine states, participants provided low DOA, $t(79) = 2.60$, $p = .03$, $d = 0.58$, or high DOA, $t(79) = 4.46$, $p < .001$, $d = 0.99$, made faster classifications than participants provided no automation. There was no difference in classification RT between participants provided high DOA and those provided low DOA, $t(78) = 2.04$, $p = .13$. For automation removal state, the simple effect test revealed no significant difference between the conditions, $F(2, 115) = 2.10$, $p = .13$, $\eta^2 = .04$. In summary, both low and high DOA benefited classification RT during routine states, and there were no return-to-manual deficits to classification RT following the use of low or high DOA.

CPA task. For CPA accuracy, there was a main effect of condition and a Condition × State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 119) = 36.17$, $p < .001$, $\eta^2 = .38$. During routine states, there was no difference in CPA accuracy between the no automation and low DOA conditions, $t < 1$. Participants provided high DOA made more accurate CPA decisions than participants provided no automation, $t(80) = 7.95$, $p < .001$, $d = 1.76$, or low DOA, $t(78) = 6.38$, $p < .001$, $d = 1.47$. For automation removal state, the simple effect test revealed a significant difference between the conditions, $F(2, 119) = 6.63$, $p = .002$, $\eta^2 = .10$. There was no difference in return-to-manual CPA accuracy between the no automation and low DOA conditions, $t < 1$. Participants previously using high DOA made *more* accurate CPA decisions than participants using no automation, $t(80) = 2.57$, $p = .04$, $d = 0.57$, or low DOA, $t(78) = 3.58$, $p = .003$, $d = 0.78$. In summary, high DOA benefited CPA accuracy during both routine states and after automation was removed.

For CPA RT, there was a main effect of condition but no Condition × State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 119) = 16.10$, $p < .001$, $\eta^2 = .22$. During routine states, participants provided low DOA made slower CPA decisions than those provided no automation, $t(78) = 3.29$, $p = .01$, $d = 0.73$, or high DOA, $t(77) = 5.46$, $p < .001$, $d = 1.22$. The difference in CPA RT for the high DOA condition compared with the no automation condition did not reach significance, $t(79) = 2.24$, $p = .08$. For automation removal state, the simple effect test revealed no significant difference between the conditions, $F(2, 115) = 1.58$, $p = .21$, $\eta^2 = .03$. In summary, low DOA impaired CPA RT during routine states, and there were no return-to-manual deficits for CPA RT following the use of low or high DOA.

Non-Automated Task Performance

For dive task accuracy, there was a Condition × State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 119) = 5.27$, $p = .01$, $\eta^2 = .08$. During routine states, participants provided low DOA, $t(80) = 3.02$, $p = .01$, $d = 0.68$, and participants provided high DOA, $t(80) = 2.71$, $p = .02$, $d = 0.60$, made poorer dive decisions than participants provided no automation. Dive accuracy was not further degraded by the use of high compared with low DOA, $t < 1$. For automation removal state, the simple effect test revealed no significant difference between the conditions, $F(2, 119) = 0.03$, $p = .97$, $\eta^2 = .00$. In summary, dive task accuracy was poorer with the use of low or high DOA, but there were no return-to-manual deficits for dive accuracy following the use of low or high DOA.

For dive task RT, there was no main effect of condition or Condition × State interaction; thus, no follow-up simple effect analyses were conducted.

Workload

For the ATWIT subjective workload measure, there was a main effect of condition and a Condition × State interaction. The simple effect test revealed a significant difference between conditions during routine states, $F(2, 117) = 4.94$, $p = .01$, $\eta^2 = .08$. During routine states, there was no significant difference in ATWIT ratings between the no automation and low DOA conditions, $t < 1$. Participants provided high DOA reported lower ATWIT ratings than participants provided no automation, $t(78) = 2.64$, $p = .02$, $d = 0.59$, or low DOA, $t(78) = 2.64$, $p = .03$, $d = 0.90$. For automation removal state, the simple effect test revealed a significant difference

between the conditions, $F(2, 119) = 4.95$, $p = .01$, $\eta^2 = .08$. After automation removal, participants previously using low DOA made higher ATWIT ratings than participants provided no automation, $t(80) = 3.23$, $p = .01$, $d = 0.71$. There was no significant difference in ATWIT ratings between the low DOA and high DOA conditions, $t < 1$, or the high DOA and no automation conditions, $t < 1$.

For NASA-TLX, there was a Condition × State interaction. However, the simple effect tests revealed no significant difference between the conditions during routine states, $F(2, 118) = 2.52$, $p = .09$, $\eta^2 = .04$, or during automation removal states, $F(2, 118) = 2.83$, $p = .06$, $\eta^2 = .05$.

In summary, high DOA reduced workload during routine states, and workload was increased after low DOA removal, but only as measured by ATWIT. Although the effects trended in the same direction (Table 3), they did not reach significance when workload was measured by the NASA-TLX.

SA

For SA, there was a Condition × State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 117) = 5.20$, $p = .01$, $\eta^2 = .08$. During routine states, participants provided low DOA made less accurate SAGAT responses than participants provided no automation, $t(78) = 3.10$, $p = .01$, $d = 0.66$. There was no difference in SAGAT accuracy between the low DOA and high DOA conditions, $t < 1$, or the high DOA and no automation conditions, $t(78) = 2.35$, $p = .07$. For automation removal state, a simple effect test on SAGAT accuracy revealed no significant difference between the conditions, $F(2, 119) = 0.50$, $p = .61$, $\eta^2 = .01$. In summary, SAGAT accuracy was poorer with the use of low DOA, and there were no return-to-manual deficits for SAGAT following the use of low or high DOA.

Discussion

The aim of Experiment 1 was to examine the impact of low DOA and high DOA on performance, workload, and SA both during routine states and after automation removal. Our predictions regarding the effects of DOA were summarized in Table 1. These predictions were based on the findings of a previous experiment in this task domain (Chen et al., 2017) and on the Onnasch et al. (2014) meta-analysis. Our findings are summarized in Table 5.

The use of low DOA in Experiment 1 benefited classification RT, but no other automated task performance metrics, compared with when no automation was provided. Furthermore, workload was not reduced with the use of low DOA, and workload increased (as measured by ATWIT) when low DOA was removed compared with no automation. There were also costs to dive task accuracy and SA with the use of low DOA compared with no automation during routine states, but these costs disappeared after the automation was removed. Other than the fact that we did not find a benefit to classification accuracy with low DOA, these findings for the low DOA condition compared with the no automation condition replicate Chen et al. (2017).

The use of high DOA benefited classification (accuracy/RT), CPA (accuracy), and lowered workload (as measured by ATWIT, but not the NASA-TLX) compared with the use of no automation. The use of high DOA also benefited classification (accuracy), CPA (accuracy/RT), and lowered workload (as measured by ATWIT) compared with the use of low DOA. The use of high DOA did not cost SA compared with no automation, but did impair dive task accuracy (but no more than when compared with low DOA). Contrary to our predictions made on the basis of the Onnasch et al. (2014) meta-analysis, high DOA removal did not cost classification/CPA performance, workload, SA, or dive task performance compared with no automation or low DOA.

EXPERIMENT 2

In Experiment 1, the use of high DOA provided several benefits to automated task performance and workload, without costs to SA, dive task performance, or return-to-manual control, when compared with the use of low DOA. It is evident that high DOA is superior to low DOA, at least in the context of simulated submarine track management, and therefore we did not further test low DOA in Experiment 2.

The removal of automation in Experiment 1 can be linked to the type of automation failure

TABLE 5: Summary of Findings From Experiment 1

Task	Routine	Matches Prediction	Removal	Matches Prediction
Classification				
Accuracy	[None = Low] < High (benefits to accuracy with high DOA)	Partial	None = Low = High (no RTM effects)	Partial
RT	None > [Low = High] (faster decisions with either low or high DOA)	Partial	None = Low = High (no RTM effects)	Partial
CPA				
Accuracy	[None = Low] < High (benefits to accuracy with high DOA)	Yes	[None = Low] < High (higher accuracy after high DOA removal)	Partial
RT	[High = None] < Low (slower decisions with low DOA)	Partial	None = Low = High (no RTM effects)	Partial
Dive				
Accuracy	None > [Low = High] (poorer accuracy with either low or high DOA)	Partial	None = Low = High (no RTM effects)	Partial
RT	None = Low = High (no difference in RT)	No	None = Low = High (no RTM effects)	Partial
Workload (ATWIT)	[None = Low] > High (reduced workload with high DOA)	Yes	[None < Low] = High (higher workload after low DOA removal)	No
Workload (NASA-TLX)	None = Low = High (no difference in workload)	Partial	None = Low = High (no RTM effects)	Partial
SA	[None > Low] = High (poorer SA with low DOA)	Partial	None = Low = High (no RTM effects)	No

Note. Routine = automation is reliable; Removal = after automation is removed; Gray shading = observed result matches predicted result; None = no automation; Low = low DOA; High = high DOA; DOA = degree of automation; RTM = return to manual; RT = response time; CPA = closest point of approach. ATWIT = Air Traffic Workload Input Technique; NASA-TLX = National Aeronautics and Space Administration Task Load Index; SA = situation awareness.

that Wickens, Clegg, et al. (2015) referred to as "automation gone," in which automation is removed. Wickens, Clegg, et al. (2015) also noted that automation may not be removed, but rather begin to provide incorrect information—a condition they referred to as "automation wrong." In comparing these two types of failures, Wickens, Clegg, et al. (2015) found that operators had more difficulty detecting and compensating for automation wrong failures than automation gone failures. Here, in Experiment 2, we aim to test whether these two types of automation failures have differential effects when individuals use high DOA.

Participants completed three scenarios, and during the last scenario, the automation was either removed (automation gone) or incorrect (automation wrong). The automation gone and automation wrong conditions were identical during routine states in that high DOA provided decision

recommendations for the classification and CPA tasks. However, the conditions differed when the automation was removed/failed. For the automation gone condition, the automation was removed as in Experiment 1. For the automation wrong condition, the automation started providing incorrect advice for the classification task. Participants from both automation conditions were instructed to report as soon as they noticed that automation was providing wrong advice.

Of particular interest was whether participants in the automation wrong condition would notice the automation failure, and if they did how long it would take them to do so. We were also interested in whether there would be any performance deficits immediately following the automation failure. Specifically, we examined performance immediately following the automation failure by analyzing classification performance (the task on which the automation was providing incorrect recommendations) on the first three classification events after the automation failure. To the extent that participants take some time to detect that the automation is providing incorrect classification recommendations, we predicted that classification accuracy on the first three events after the automation failure could be poorer, and classification RT slower, for the automation wrong condition compared with the no automation condition. In contrast, we did not expect to see classification accuracy or RT deficits immediately following the automation failure for the automation gone condition compared with the no automation condition because participants were notified that automation was no longer available, and the evidence to date from Chen et al. (2017) and the current Experiment 1 suggests that participants should be able to regain manual control relatively quickly under these circumstances.

In addition to these aforementioned novel analyses, we expected to replicate the benefits to automated task performance and workload, and deficits to the dive task, for the two high DOA conditions compared with the no automation condition during routine states. The importance of establishing the robustness of psychological effects has received much recent attention (e.g., Pashler & Wagenmakers, 2012) and is particularly vital when the resulting knowledge could be used by practitioners in safety-critical work settings (Jones, Derby, & Schmidlin, 2010). The replication is also important because unlike Experiment 1, in Experiment 2 participants who were using high DOA were instructed that although automation was highly reliable, it may not be perfect, which may decrease the extent to which they trust and rely on the automated recommendations (see Lee & See, 2004). Automation removal state in Experiment 2 was defined as the time after automation was removed for the automation gone condition (as in Experiment 1), and as the time period after participants detected the automation failure for the automation wrong condition. Note that correctly reporting the automation failure resulted in the automation being disengaged (but participants were not informed that this would occur). Based on Experiment 1 results, we did not expect return-to-manual deficits during the automation removal state for the high DOA conditions compared with the no automation condition.

An additional goal of Experiment 2 was to examine how participants rated the importance of each of the three tasks. In Experiment 1, we found significant costs to the dive task that replicated those reported by Chen et al. (2017). We wanted to investigate the possibility that the dive task deficit was due to participants placing less importance on their performance on the dive task compared with the other two tasks due to the fact that the dive task was the only task that was not automated.

Method

Participants. Participants were 120 (70 females) undergraduate psychology students (age: M = 21.7, SD = 6.17) who participated for course credit and were randomly assigned to one of three conditions: no automation (n = 40), automation gone (n = 40), and automation wrong (n = 40).

Simulated submarine track management task. The simulation was identical to the no automation and high DOA conditions from Experiment 1 with the following exceptions. For the automation gone and automation wrong conditions, the automation was unexpectedly removed or failed during the last scenario at 10.38, 10.48, or 10.88 min into the 27.5-min

scenario. The message provided to participants in the automation gone condition was identical to that used for the high DOA condition in Experiment 1. In the automation wrong condition, the automation started providing incorrect advice for the classification task. Specifically, the horizontal lines were placed either 30 s too early or late on the soundtracks, and the recommended classification was incorrect (e.g., if the contact was an enemy, the classification letter presented next to the line was f, t, or m).

The Track Assist interface was modified to include a "fail" button. This button was available from the beginning of each scenario. Participants (in both automation conditions) were instructed to click this button if they believed the automation was providing wrong advice. When clicked, a message appeared saying "Automation failure detected. Track Assist turned off. Manual tracking required" and participants had to acknowledge that they had read this message by clicking the "ok" button. If the fail button was clicked when the automation was functioning correctly, a message appeared saying "Automation has not failed" and the automation continued operating as usual.

Measures and procedure. Workload and SA measures were identical to Experiment 1. Participants rated the perceived importance of each task on a 5-point Likert-type scale (1 = *not at all important* to 5 = *extremely important*) after the last scenario. The training was the same as in Experiment 1, but there was an additional instruction specifying that although the automation was highly reliable, it may not be perfect, and participants were instructed how to report an automation failure. Each participant completed three scenarios in their assigned condition, and the order of scenarios was counterbalanced.

Results

The mean RT to detect the automation failure by participants in the automation wrong condition was 174.13 s; 95% CI = [125.64, 222.61]. As seen in Figure 2, 50% of the participants in the automation wrong condition had not reported the automation failure 173.87 s after the failure occurred. Three participants in the automation wrong condition did not detect the automation failure at all.

Classification Task Performance Immediately Following the Automation Failure

We analyzed performance on the first three classification events after the automation failure. The classification accuracy and RT data for these three classification events are presented in Figure 3. To test our predictions, we ran 3 Condition (no automation, automation gone, automation wrong) × 3 Classification Event (first event after failure, second event after failure, third event after failure) mixed ANOVAs on classification accuracy and on classification RT. The between-subjects factor was condition and the within-subjects factor was classification event. We planned to follow-up significant main effects of condition, or interactions between condition and classification event, with tests of simple effects separately (with Bonferroni corrections), comparing the three conditions on each classification event.

For classification accuracy, it was predicted that performance immediately after the automation failure would be poorer for the automation wrong condition compared with the no automation condition, but there would be no difference in performance between the automation gone and no automation conditions. A mixed ANOVA on classification accuracy revealed a main effect of classification event, $F(2, 234) = 3.21$, $p = .04$, $\eta_p^2 = .03$, but no main effect of condition, $F(1, 117) = 0.55$, $p = .587$, $\eta_p^2 = .01$, and no interaction effect $F(2, 234) = 2.24$, $p = .07$, $\eta_p^2 = .04$.

For classification RT, it was predicted that RT would be slower immediately after the automation failure for the automation wrong condition compared with the no automation condition, but that there would be no difference in RT between the automation gone and no automation conditions. A mixed ANOVA on classification RT revealed no main effect of classification event, $F(2, 128) = 1.62$, $p = .20$, $\eta_p^2 = .03$; no main effect of condition, $F(1, 64) = 2.09$, $p = .13$, $\eta_p^2 = .06$; and no interaction effect, $F(2, 128) = 0.98$, $p = .42$, $\eta_p^2 = .03$.

In brief, there were no reliable differences in classification accuracy or RT between the conditions for the three classification events after

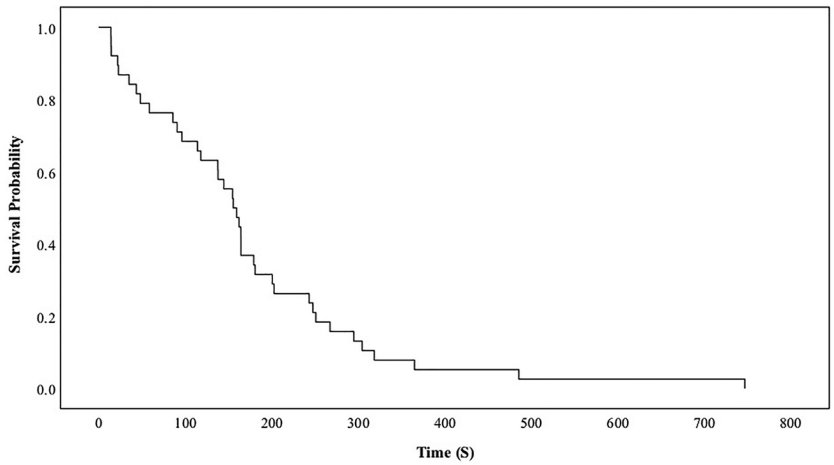


Figure 2. A Kaplan–Meier survival analysis representing the time (in seconds) taken for participants in the automation wrong condition to detect the automation wrong failure.

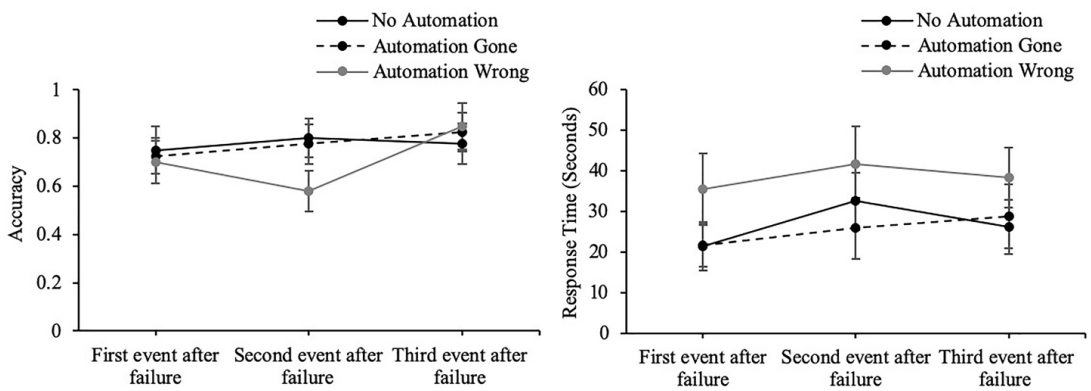


Figure 3. Classification accuracy (left graph) and RT (right graph) for the first three classification events after the automation failure, as a function of automation condition. Error bars represent 95% between-subjects confidence intervals.

the automation failure. Thus, performance immediately after the automation failure was not poorer for the automation wrong condition compared with the no automation condition. As predicted, there was no difference in performance between the automation gone and no automation conditions.

Routine and Automation Removal States

We then conducted analyses to replicate the results from Experiment 1. The data were separated into the time period automation was available (routine state: first two scenarios and one third of the last scenario) and time period automation was removed (automation removal: two thirds of the last scenario for the automation gone condition, and from the time point, automation was turned off by the participant for the automation wrong condition). The three participants in the automation wrong condition who never reported the automation failure were excluded from the analyses reported in the following text.

TABLE 6: Descriptive Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 2

	Classification		CPA		Dive		SAGAT	Workload Rating	
Automation	Hit	RT	Hit-FA	RT	Hit-FA	RT	Accuracy	ATWIT	NASA-TLX
Routine state									
None	0.73 [0.65, 0.81]	30.82 [26.37, 35.28]	0.38 [0.30, 0.46]	21.09 [15.27, 26.92]	0.72 [0.67, 0.77]	9.88 [9.09, 10.66]	0.53 [0.48, 0.58]	4.98 [4.57, 5.39]	61.54 [57.87, 65.21]
High	0.94 [0.90, 0.96]	21.41 [19.91, 22.92]	0.80 [0.74, 0.86]	11.97 [10.44, 13.50]	0.59 [0.55, 0.64]	9.46 [8.73, 10.19]	0.51 [0.49, 0.54]	4.11 [3.81, 4.40]	52.09 [48.79, 55.38]
Automation removal state									
None	0.72 [0.62, 0.82]	28.61 [23.37, 34.05]	0.28 [0.19, 0.36]	22.18 [15.63, 28.73]	0.77 [0.69, 0.85]	9.88 [7.93, 11.84]	0.54 [0.49, 0.59]	5.19 [4.77, 5.61]	61.66 [57.73, 66.78]
High	0.77 [0.72, 0.82]	31.80 [28.26, 35.34]	0.36 [0.31, 0.41]	19.95 [14.54, 25.36]	0.77 [0.72, 0.83]	9.60 [8.61, 10.59]	0.53 [0.50, 0.57]	5.50 [5.15, 5.86]	58.06 [54.62, 61.51]

Note. The 95% between-subjects confidence intervals are presented in brackets. CPA = closest point of approach; SAGAT = Situation Awareness Global Assessment Technique; RT = response time; FA = false alarm; ATWIT = Air Traffic Workload Input Technique; NASA-TLX = National Aeronautics and Space Administration Task Load Index.

The automation gone and automation wrong conditions were identical during routine states in that they provided high DOA for the classification and CPA tasks, and identical at automation removal in that all participants knew the automation was no longer available (removed, or detected as having failed and removed). On this basis, we combined the data from the two high DOA conditions from Experiment 2. The means and between-subjects 95% CIs for performance, workload, and SA are presented in Table 6.

We compared the two high DOA conditions to the no automation condition with 2 Condition (high DOA, no automation) × 2 Automation State (routine, automation removal) mixed ANOVAs with the between-subjects factor as condition and the within-subjects factor as automation state. The inferential statistics from the ANOVAs are summarized in Table 7. Significant main effects of condition, and interactions between condition and automation state, were followed by comparisons of simple effects (*t* tests) conducted separately for the routine state and the automation removal state, and are presented in text. We had a power of 0.82 to detect a medium-to-large effect size (Cohen, 1988). As in Experiment 1, several main effects of automation state were found. They are reported in Table 7 but for brevity are not further discussed.

Automated Task Performance

Classification task. For classification accuracy, there was a main effect of condition and a Condition × State interaction. During routine states, participants provided high DOA made more accurate contact classifications compared with participants provided no automation, $t(118) = 5.89$, $p < .001$, $d = 1.02$. After automation removal, there was no difference in classification accuracy between the conditions, $t < 1$. For classification RT, there was a Condition × State interaction. During routine states, participants provided high DOA made faster classifications than those provided no automation, $t(118) = 4.99$, $p < .001$, $d = 0.86$. After automation removal, there was no difference in classification RT between the conditions, $t < 1$. These findings for the classification task replicate Experiment 1.

CPA task. For CPA accuracy, there was a main effect of condition and a Condition × State

TABLE 7: Inferential Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 2

	High Degree of Automation vs. No Automation				
Dependent Variable	Effect	*F*	*df*	*p*	η_p^2
Classification (Hit)	Condition	11.24	(1, 115)	.001*	.09
	State	33.50	(1, 115)	<.001*	.23
	Interaction	23.99	(1, 115)	<.001*	.17
Classification (RT)	Condition	2.22	(1, 112)	.14	.02
	State	6.34	(1, 112)	.02*	.05
	Interaction	18.21	(1, 112)	<.001*	.14
CPA (Hit-FA)	Condition	37.87	(1, 115)	<.001*	.25
	State	111.30	(1, 115)	<.001*	.49
	Interaction	43.20	(1, 115)	<.001*	.27
CPA (RT)	Condition	3.63	(1, 112)	.06	.03
	State	5.24	(1, 112)	.02*	.05
	Interaction	2.31	(1, 112)	.13	.02
Dive (Hit-FA)	Condition	2.55	(1, 115)	.11	.02
	State	39.40	(1, 115)	<.001*	.26
	Interaction	11.11	(1, 115)	.001*	.09
Dive (RT)	Condition	0.55	(1, 114)	.46	.01
	State	0.03	(1, 114)	.87	<.001
	Interaction	0.02	(1, 114)	.88	<.001
NASA-TLX	Condition	6.37	(1, 118)	.01*	.05
	State	7.00	(1, 118)	.01*	.06
	Interaction	6.46	(1, 118)	.01*	.05
ATWIT	Condition	1.13	(1, 115)	.29	.01
	State	55.29	(1, 115)	<.001*	.33
	Interaction	29.65	(1, 115)	<.001*	.21
SAGAT (Accuracy)	Condition	0.19	(1, 115)	.67	.002
	State	0.50	(1, 115)	.48	.004
	Interaction	1.00	(1, 115)	.76	.001

Note. RT = response time; CPA = closest point of approach; FA = false alarm; NASA-TLX = National Aeronautics and Space Administration Task Load Index; ATWIT = Air Traffic Workload Input Technique; SAGAT = Situation Awareness Global Assessment Technique.
*$p < .05$.

interaction. During routine states, participants provided high DOA made more accurate CPA task decisions compared with participants provided no automation, $t(118) = 8.28$, $p < .001$, $d = 1.63$. After automation removal, there was no difference in CPA accuracy between the conditions, $t < 1$. For CPA RT, there was no main effect of condition or Condition × State interaction. These findings for the CPA task replicate Experiment 1.

Non-Automated Task Performance

For dive task accuracy, there was a Condition × State interaction. During routine states, participants provided high DOA made less accurate dive task decisions compared with participants provided no automation, $t(118) = 3.37$, $p = .001$, $d = 0.68$. After automation removal, there was no difference in dive accuracy between the conditions, $t < 1$. For dive RT, there was no main

TABLE 8: Descriptive Statistics for Task Importance Ratings by Condition in Experiment 2

Automation	Classification	CPA	Dive
None	4.20 [3.92, 4.48]	2.98 [2.65, 3.30]	3.75 [3.39, 4.11]
Gone	4.18 [3.89, 4.46]	3.08 [2.75, 3.40]	3.60 [3.26, 3.94]
Wrong	4.15 [3.86, 4.44]	3.05 [2.73, 3.37]	3.48 [3.11, 3.84]

Note. The 95% between-subjects confidence intervals are presented in brackets. CPA = closest point of approach.

effect of condition or Condition × State interaction. These findings for the dive task replicate Experiment 1.

Workload

For ATWIT, there was a Condition × State interaction. During routine states, participants provided high DOA made lower ATWIT ratings compared with participants provided no automation, $t(118) = 3.44$, $p = .001$, $d = 0.67$. After automation removal, there was no difference in ATWIT ratings between the conditions, $t < 1$. These findings for the ATWIT replicate Experiment 1. In addition, for NASA-TLX, the main effect of condition and Condition × State interaction reached significance. During routine states, participants provided high DOA made lower NASA-TLX ratings compared with participants provided no automation, $t(118) = 3.54$, $p = .001$, $d = 0.71$. After automation removal, there was no difference in NASA-TLX ratings between the conditions, $t < 1$.

SA

There was no main effect of condition or Condition × State interaction, replicating Experiment 1.

Task Importance

The task importance ratings are presented in Table 8. A 3 Condition (no automation, automation gone, automation wrong) × 3 Task Type (classification, CPA, dive) mixed ANOVA on task importance ratings revealed a main effect of task type, $F(2, 234) = 47.22$, $p < .001$, $\eta_p^2 = .29$. Participants rated the classification task as being more important than the CPA task, $t(119) = 10.57$, $p < .001$, $d = 1.20$, and the dive task, $t(119) = 5.20$, $p < .001$, $d = 0.56$. In addition, the dive task was rated as being more important than the CPA task, $t(119) = 4.35$, $p < .001$, $d = 0.55$. There was no main effect of condition and no interaction. Thus, there were no differences in dive task importance ratings between the conditions, suggesting that participants in the automated conditions did not place less importance on their performance on the dive task compared with participants not provided automation.

GENERAL DISCUSSION

In Experiment 1, we examined the effects of low and high DOA on performance, workload, SA, non-automated task performance, and return-to-manual performance. Low DOA provided information acquisition and analysis support. High DOA provided decision recommendation support while still requiring participants to execute task actions. Participants completed two tasks that were supported by the automation (classification and CPA), and one task that was not supported by automation (dive). In Experiment 2, when automation failed, it was either removed completely (automation gone condition), as in Experiment 1, or started providing incorrect advice for the classification task (automation wrong condition). We examined whether participants would notice the automation wrong failure and if so, how long it would take them to do so. We also examined whether there would be any performance deficits immediately following the automation failure on the classification task. Furthermore, in Experiment 2, we expected to replicate the benefits to automated task performance and workload, and the costs to non-automated task performance, with the use of high DOA that were found in Experiment 1.

The Benefits and Costs of Low and High DOA

We found little evidence of a benefit in using low DOA. Only one of the four automated task performance metrics (classification

RT) improved, and there was no reduction in workload. In addition, there were costs to dive task accuracy and SA with the use of low DOA compared with no automation during routine states. These findings largely replicate those reported by Chen et al. (2017).

Compared with the use of low DOA, the use of high DOA benefited three automated task performance metrics (classification accuracy, and CPA accuracy/RT). Participants also reported lower workload with the use of high DOA compared with low DOA and no automation. In addition, although high DOA did cost non-automated task performance compared with no automation, the extent of this cost was not larger than that for the low DOA condition. Also, in Experiments 1 and 2, after the automation was no longer available (removed, or detected as having failed and removed), it was not more difficult for participants previously using high DOA to regain manual control. Therefore, participants were able to effectively return-to-manual control after knowing that automation was no longer available. Overall, we have found evidence that under some conditions, it is possible that moving from a low DOA to a high DOA can provide a "free lunch," that is, it can increase the benefits of automation without further increasing the costs (see Wickens, 2018).

At first glance, our findings of increased benefits without further costs when using high compared with low DOA seem inconsistent with the Onnasch et al. (2014) meta-analytic finding that the negative consequences of automation are more likely, the higher the DOA. However, close inspection of Onnasch et al. indicates that their meta-analysis contained substantial variance in effect size between studies, with trends and effects ranging from strong to weak or even reversed. This variance is likely due to the variability in the nature of the tasks used across studies included in the Onnasch et al. meta-analysis. In addition, many of the studies in the Onnasch et al. meta-analysis used relatively fast evolving tasks such as air traffic control and unmanned vehicle control. In contrast, a key feature of submarine track management is the very slow pace in which contacts move on the display. High DOA may have indeed reduced the extent to which participants actively processed raw information (e.g., contact position and heading) (complacency; Parasuraman & Manzey, 2010), but the effect of this may have been attenuated by the slow pace of the task that allowed sufficient time for participants to recover. Furthermore, Onnasch et al. suggest that the negative consequences of automation were the strongest when DOA moved from supporting information acquisition and analysis to also supporting action selection/action execution. It is worth noting that in the current study, high DOA did not cross the boundary between decision recommendation and action execution because participants were still required to execute the final task action.

The benefits to automated task performance and workload for the high compared with low DOA and no automation conditions (Experiment 1), and for the high DOA conditions compared with the no automation condition (Experiment 2), were reasonably consistent. In Experiments 1 and 2, there were also clear and consistent costs to dive task accuracy during routine states with the use of low and high DOA compared with the no automation condition. Even with the reduction in workload with the use of high DOA compared with no automation, performance on the dive task degraded. It would have been reasonable to expect that the reduced workload associated with the use of high compared with no automation should have provided the operator with the additional cognitive capacity to more effectively manage the non-automated dive task (Manzey et al., 2012; Rovira, McGarry, & Parasuraman, 2007). Nevertheless, reduced workload with high DOA would only have benefited dive task performance to the extent that the spared capacity was directly allocated toward scrutinizing contact location and heading information. It seems that participants who were provided with high DOA for classification and CPA tasks scrutinized contact location and heading information less closely than participants who were not provided with automation (complacency), and the dive task likely suffered because it also required assessment of contact location/heading information. To further test this explanation, future research could examine performance on a non-automated task that is independent of the automated tasks. Performance on an independent

non-automated task should be the same or if not better for those who receive high DOA compared with those who receive no automation, due to the spare cognitive capacity from the reduction in workload with the use of automation. Note that in Experiment 2, we ruled out the possibility that the dive task deficit could be due to participants provided with automation placing less importance on the dive task compared with the two other automated tasks.

In Experiment 2, we predicted that classification performance immediately after the high DOA failure would be poorer for the automation wrong condition compared with the no automation condition. However, this prediction was not supported. Interestingly, although it took participants on average 3 min to report the automation wrong failure, they were still able to correctly classify contacts immediately after the failure as successfully as the no automation condition. In post hoc analyses, at each of the three classification events immediately after the failure, we split participants in the automation wrong condition according to whether they had reported the automation failure or not. There was still no significant difference in classification accuracy or RT on the three classification events immediately after the automation failure for the subset of participants who had not yet reported the failure, compared with the no automation condition or automation gone condition. This suggests that participants may have become suspicious about the accuracy of the automation and started to make their own manual classification decisions, but decided to allow some time to clarify and ensure that the automation was not performing accurately before they formally reported the failure.

PRACTICAL IMPLICATIONS AND CONCLUSIONS

A key question in complex work systems is to what extent decision recommendation automation can be effectively used. The results of the current study suggest that automation that recommends decisions can be effectively used and, in the current context, was superior to a low DOA that only provided information acquisition and analysis support. Specifically, automation that recommended decisions leads to performance and workload benefits without any costs to SA or return-to-manual performance, compared with automation that provided information acquisition and analysis support. Although the current study used a simulation of submarine track management, the findings of this work are also relevant to other work contexts, particularly those involving slowly evolving contexts that require operators to monitor demanding perceptual displays (e.g., maritime surveillance).

Automating tasks can improve operator performance and reduce workload, but accidents have occurred because human operators have been unprepared to take over when manual control is required. If automation fails, the operator's ability to resume manual control is critical. In the current study, although participants were able to regain manual control, it took on average 3 min for them to detect that automation was providing incorrect advice. As discussed, the fact there was no decrement to classification performance immediately following the automation failure suggests at least some participants who had not indicated that there was a failure were suspicious that automation may have not been performing accurately and were making manual classification decisions. Nonetheless, in a context where there is more time pressure (e.g., unmanned vehicle control, air traffic control) to make a manual decision, such a delay could be catastrophic. Future research could examine whether operators would still take as long to formally register an automation wrong failure in a faster updating task. Before implementing automation that recommends decisions, designers should carefully consider the level of uncertainty in the environment (i.e., the chance that automation may be incorrect) and the operational consequences of a loss of SA or return-to-manual performance deficits.

The simulated submarine track management task used in the current study was designed based on a task analysis conducted with Royal Australian Navy Submariners. Accordingly, the current experiments have external validity as they represent a typical example of a work context that requires operators to monitor demanding perceptual displays. That said, we are aware of the potential problems in generalizing from novice participants to expert operators as there

are undoubtedly differences in their cognitive skills and motivation. Future research could examine how expert submariners are affected differently by DOA and task type in the current simulated submarine track management task. There is, however, evidence to suggest that our results with novice participants can validly inform practical issues in operational contexts. A study by Loft et al. (2016) found relatively consistent results across novice participants using the current simulated submarine track management task and expert submariners using real submarine combat systems. In addition, Onnasch et al. (2014) found that expertise did not moderate the benefits and costs of automation; thus, benefits and costs were as statistically likely to occur for experts as they were for novice participants.

In conclusion, the automated system that recommended decisions was effectively utilized by participants in the current context and appeared to be superior to the automated system that supported information acquisition and analysis. Automation that recommends decisions is appropriate in contexts where the consequences of an automation failure are not serious enough to outweigh the benefits. However, designers should be cautious and consider the level of uncertainty in the environment and the consequences of a loss of SA or return-to-manual performance deficits before implementing decision-aiding automation. In contexts where return to manual performance is of serious concern, operators should be kept involved in the action selection and execution stages.

ACKNOWLEDGMENT

This research was supported by Discovery Grant DP160100575 awarded to Loft from the Australian Research Council.

KEY POINTS

- With the ongoing emphasis on developing high degree of automation (DOA) that can recommend decisions to operators, it is critical to further understand how high DOA systems affect the human operator.
- In a simulated submarine track management task, high DOA that provided decision recommendations provided benefits to performance and workload, without additional costs to SA or non-automated task performance, compared with low DOA.
- There were no return-to-manual deficits when participants had knowledge that low DOA or high DOA was disengaged.
- Participants using high DOA took on average 3 min to notice that automation was providing incorrect recommendations, but there was no deficit to performance immediately following the automation failure.
- Designers should consider the level of uncertainty in the environment and the consequences of a loss of situation awareness (SA) or return-to-manual deficits before implementing decision-aiding automation.

ORCID iDS

Monica Tatasciore https://orcid.org/0000-0001-7290-0225

Troy A. W. Visser https://orcid.org/0000-0003-3960-2263

Steph I. C. Michailovs https://orcid.org/0000-0002-6767-6692

REFERENCES

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*, 1065–1105. doi:10.1177/0149206309352880

Chen, S. I., Visser, T. A. W., Huf, S., & Loft, S. (2017). Optimizing the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied*, *23*, 240–262. doi:10.1037/xap0000126

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *32*, 97–101.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*, 65–84. doi:10.1518/001872095779049499

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, *59*, 5–27. doi:10.1177/0018720816681350

Hart, S. G., & Staveland, L. E. (1987). Development of NASA-TLX: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, The Netherlands: Elsevier.

Jones, K. S., Derby, P. L., & Schmidlin, E. A. (2010). An investigation of the prevalence of replication research in human factors. *Human Factors*, *52*, 586–595. doi:10.1177/0018720810384394

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*, 113–153. doi:10.1080/1463922021000054335

Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *10*, 409–430.

Kirschenbaum, S. S. (2011). Expertise in the submarine domain: The impact of explicit display on the interpretation of uncertainty. In K. L. Mosier & U. M. Fischer (Eds.), *Informed by knowledge: Expert performance in complex situations* (pp. 189–199). New York, NY: Psychology Press.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80. doi:10.1518/hfes.46.1.50_30392

Li, H., Wickens, C. D., Sarter, N., & Sebok, A. (2014). Stages and levels of automation in support of space teleoperations. *Human Factors*, *56*, 1050–1061. doi:10.1177/0018720814522830

Loft, S., Morrell, D. B., Ponton, K., Braithwaite, J., Bowden, V., & Huf, S. (2016). The impact of uncertain contact location on situation awareness and performance in simulated submarine track management. *Human Factors*, *58*, 1052–1068. doi:10.1177/0018720816652754

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*, 57–87. doi:10.1177/1555343411433844

Merlo, J. L., Wickens, C. D., & Yeh, M. (2000). Effect of reliability on cue effectiveness and display signaling. In *Proceedings of the 4th Annual Army Federated Laboratory Symposium* (pp. 27-31). College Park, MD: Army Research Federated Laboratory Consortium.

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. doi:10.1177/0018720813501549

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*, 381–410. doi:10.1177/0018720810376055

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation induced "complacency." *The International Journal of Aviation Psychology*, *3*, 1–23. doi:10.1207/s15327108ijap0301_1

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253. doi:10.1518/001872097778543886

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

Roberts, A., Stanton, N. A., & Fay, D. (2017). The command team experimental test-bed phase two: Assessing cognitive load and situation awareness in a submarine control room. In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation* (pp. 427–437). Berlin, Germany: Springer.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76–87.

Sheridan, T. B. (2015). Automation. In D. A. Boehm-Davis, F. T. Durso, & D. L. John (Eds.), *APA handbook of human systems integration* (pp. 449–465). Washington, DC: American Psychological Association.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). Cambridge: Man-Machine Systems Laboratory, Massachusetts Institute of Technology.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604.

Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe*. Atlantic City, NJ: Federal Aviation Administration.

U.S. Air Force. (2015). *Autonomous horizons*. Washington, DC: U.S. Air Force Office of the Chief Scientist.

Vagia, M., Transeth, A. A., & Fjerdingen, S. A. (2016). A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics*, *53*, 190–202. doi:10.1016/j.apergo.2015.09.013

Wickens, C. D. (2018). Automation stages & levels, 20 years after. *Journal of Cognitive Engineering and Decision Making*, *12*, 35–41. doi:10.1177/1555343417727438

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*, *57*, 728–739. doi:10.1177/0018720815581940

Wickens, C. D., Sebok, A., Li, H., Sarter, N., & Gacy, A. M. (2015). Using modeling and simulation to predict operator performance and automation-induced complacency with robotic automation: A case study and empirical validation. *Human Factors*, *57*, 959–975. doi:10.1177/0018720814566454

Monica Tatasciore is a master's student enrolled in the Doctor of Philosophy and Master of Industrial and Organizational Psychology program at the University of Western Australia.

Vanessa K. Bowden is a lecturer at the University of Western Australia. She received her PhD in psychology in 2012 from the University of Western Australia.

Troy A. W. Visser is an associate professor at the University of Western Australia. He received his PhD in cognitive systems in 2001 from the University of British Columbia.

Steph I. C. Michailovs is a postdoctoral research fellow at the University of Western Australia. She received her PhD in psychology in 2019 from the University of Western Australia.

Shayne Loft is an associate professor at the University of Western Australia. He received his PhD in psychology in 2004 from the University of Queensland.

Date received: November 13, 2018
Date accepted: July 3, 2019

Touchscreens for Aircraft Navigation Tasks: Comparing Accuracy and Throughput of Three Flight Deck Interfaces Using Fitts' Law

Nout C. M. van Zon, Clark Borst, Daan M. Pool, and **Marinus M. van Paassen**, Delft University of Technology, The Netherlands

Objective: Use Fitts' law to compare accuracy and throughput of three flight deck interfaces for navigation.

Background: Industry is proposing touch-based solutions to modernize the flight management system. However, research evaluating touchscreen effectiveness for navigation tasks in terms of accuracy and throughput on the flight deck is lacking.

Method: An experiment was conducted with 14 participants in a flight simulator, aimed at creating Fitts' law accuracy and throughput models of three different flight deck interfaces used for navigation: the mode control panel, control display unit, and a touch-based navigation display. The former two constitute the conventional interface between the pilot and the flight management system, and the latter represents the industry-proposed solution for the future.

Results: Results indicate less accurate performance with the touchscreen navigation display compared to the other two interfaces and the throughput was lowest with the mode control panel. The control display unit was better in both accuracy and throughput, which is found to be largely attributed to the tactile and physical nature of the interface.

Conclusion: Although performance in terms of accuracy and throughput was better with the control display unit, a question remains whether, when used during a more realistic navigation task, performance is still better compared to a touch-based interface.

Application: This paper complements previous studies in the usage of aircraft touchscreens with new empirical insights into their accuracy and throughput, compared to conventional flight deck interfaces, using Fitts' law.

Keywords: touchscreens, interface evaluation, human performance modeling, coordinated action, flight displays, Fitts' law

Address correspondence to Daan M. Pool, Control & Simulation, Department Control & Operations, Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, Zuid-Holland, 2629 HS Delft, The Netherlands; e-mail: d.m.pool@tudelft.nl.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 897–908
DOI: 10.1177/0018720819862146

INTRODUCTION

The modern-day flight management system (FMS) was introduced on the Boeing 767 in 1982 (Bulfer, 1991) to assist pilots in both lateral navigation (LNAV) and vertical navigation (VNAV). As an interface to the FMS, the control display unit (CDU) was introduced and remains the industry standard to date. For example, when the CDU on the Boeing 787 was replaced with a digital copy, the look and feel remained the same.

However, looking ahead at future developments in LNAV procedures, the necessity to modernize the FMS interface becomes evident. The SESAR Joint Undertaking expects the number of flights in European airspace to have increased by 52% in 2035 compared to 2012 (SESAR Joint Undertaking, 2014). As a result, Huisman, Verhoeven, Van Houten, and Flohr (1997) expect an increased frequency of en-route route adjustments. Van Marwijk, Borst, Mulder, Mulder, and Van Paassen (2011) call for "a redesign of the navigation planning interface [due to] increasing punctuality in, [amongst others,] European SESAR concepts, [which will] make airborne flight plan amendment increasingly complex."

Touchscreens have the potential to reduce cognitive workload and increase situation awareness due to their "intuitive" way of interaction and their flexibility in displaying additional task-relevant information, respectively (Dodd et al., 2014; Hutchins, Hollan, & Norman, 1985; Kaminani, 2011; Rogers, Fisk, McLaughlin, & Pak, 2005; Shneiderman, 1982). As such, aircraft and equipment manufacturers have been proposing touchscreens on their newest flight decks in anticipation of increased complexity in future navigation tasks. However, concerns have been voiced about the loss of tactile feedback, usability in dynamic environments

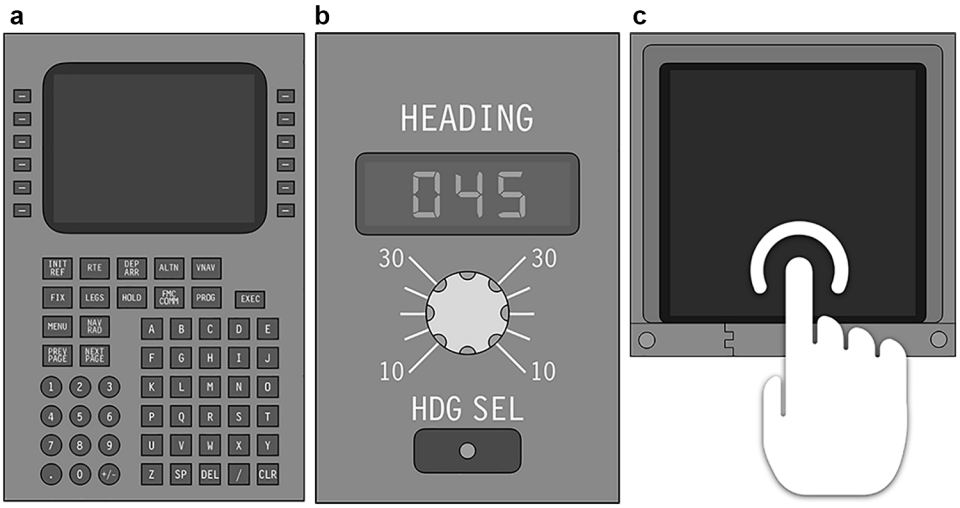


Figure 1. Three flight deck interfaces that are to be investigated: (a) heading control knob on the mode control panel (MCP), (b) control display unit (CDU), and (c) touch-based navigation display (TND).

(e.g., turbulence), and physical fatigue of operation (Degani, Palmer, & Bauersfeld, 1992; Dodd et al., 2014; Kaminani, 2011; Stuyven, Damveld, & Borst, 2012).

Previous research has been done evaluating touchscreen interfaces in general and comparing them to less direct interfaces such as trackballs, trackpads, and rotary controllers (Ballas, Heitmeyer, & Pérez-Quiñones, 1992; Bjørneseth, Dunlop, & Hornecker, 2012; Degani et al., 1992; Forlines, Wigdor, Shen, & Balakrishnan, 2007; Stanton, Harvey, Plant, & Bolton, 2013). In the aviation domain, Dodd et al. (2014) found increased task execution time, error rates, and subjective workload for touchscreen usage in turbulence and at specific cockpit positions. However, a truly comparative study between a touchscreen and conventional flight deck interfaces on a fundamental input level, quantified in terms of *input accuracy* and *information throughput* as a function of task complexity, has not yet been carried out.

The goal of this research is to develop and compare accuracy and throughput models of three flight deck interfaces used during LNAV. These interfaces are the mode control panel (MCP), the CDU, and a touch-based navigation display (TND), illustrated in Figure 1. The models will be developed based on variations of Fitts' law (Fitts, 1954). This law, first published in 1954, has been used by human–machine interaction researchers for analysis of the speed-accuracy trade-off and movement time (MT) in rapid aimed movement tasks (Jagacinski & Fisch, 1997; Jagacinski, Repperger, Ward, & Moran, 1980; Stoelen & Akin, 2010; Trudeau, Udtamadilok, Karlson, & Dennerlein, 2012), and as a valuable tool for human–machine interface design (Flach, Hagen, O'Brien, & Olson, 1990; Francis & Oxtoby, 2006; Gao & Sun, 2015; Jax, Rosenbaum, Vaughan, & Meulenbroek, 2003; MacKenzie, 1992; Soukoreff & MacKenzie, 2004). Fitts' law models also enable quantitative comparison of the effectiveness of different interfaces based on their *throughput* (Jagacinski & Fisch, 1997; MacKenzie, 1992; Soukoreff & MacKenzie, 2004), describing how many *bits* of task difficulty, as defined by an index of difficulty (ID), an interface can handle per second.

Fitts' Law

The complete and original Fitts' law model (Fitts, 1954) that describes MT as a function of ID in a high-accuracy pointing task is presented in Equation 1. Here, a and b are empirical linear regression constants, A is the amplitude (distance to be traversed), and W_e is the

effective width of the target. The latter is empirically calculated using the standard deviation of measured endpoint coordinates (Soukoreff & MacKenzie, 2004).

$$MT = a + b \cdot ID = a + b\log_2\left(\frac{A}{W_e} + 1\right)[\text{seconds}] \quad (1)$$

The usefulness of Fitts' law in this study is twofold. First, it can help build models of task execution time for a particular interface. Second, it can provide a quantitative description of the FMS interface by comparing the throughput (TP) of individual interfaces. Equation 2 defines the throughput in bits per second, which is calculated by dividing the ID by the measured MT for each participant and experimental condition. The total numbers of conditions and participants are defined by x and y, respectively. $ID_{e_{ij}}$ defines the index of difficulty, adjusted using the effective width W_e, and MT_{ij} the movement time, both for a specific experimental condition and participant.

$$TP(\text{Throughput}) = \frac{1}{y}\sum_{i=1}^{y}\left(\frac{1}{x}\sum_{j=1}^{x}\frac{ID_{e_{ij}}}{MT_{ij}}\right)[\text{bits/s}] \quad (2)$$

Mode Control Panel (MCP)

The MCP is the standard interface between the pilot and the autopilot and uses, among others, a rotary heading control knob with which the horizontal flight direction (i.e., heading) can be changed. Research by Stoelen and Akin (2010) has shown that Fitts' law can be extended to rotational input tasks by replacing the linear width and amplitude with an angular width ω and amplitude α, respectively. The effective angular width ω_e can be calculated based on the standard deviation in endpoints σ_ϕ. Stoelen and Akin (2010) found a good model fit for a smooth, continuous rotational task. The heading control knob, however, uses detents, resulting in discrete motion inputs, and there is no literature regarding its effectiveness in terms of Fitts' law. Despite potential inappropriateness of the model proposed by Stoelen and Akin (2010), shown in Equation 3, this research still used it in modeling the heading control knob on the MCP.

$$MT = a + b\log_2\left(\frac{\alpha}{\omega_e} + 1\right) \text{ where } \omega_e = \sqrt{2\pi e}\sigma_\phi \quad (3)$$

Control Display Unit (CDU)

The CDU is a keyboard-type input device by which pilots can change a planned flight route by entering or deleting waypoints. Research by MacKenzie and Buxton (1992) and Soukoreff and MacKenzie (1995) has shown that Fitts' law can be extended to keyboard data-entry tasks. The model, shown in Equation 4, is based on an assumption that using either the minimum height H or width W of the target in the computation of the ID is sufficient. MacKenzie and Buxton (1992) have found this to provide adequate results. In the case of a key-repeat task, the amplitude is zero and thus the ID, namely $\log_2(0+1)$, will equal zero. Therefore, Soukoreff and MacKenzie (1995) propose an averaged repeat movement time parameter MT_{repeat} for such tasks.

$$MT_{ij} = \begin{cases} a + b\log_2\left(\dfrac{A_{ij} + \min(H_j, W_j)}{\min(H_j, W_j)}\right) & \text{if } i \neq j \\ MT_{repeat} & \text{if } i = j \end{cases} \quad (4)$$

Furthermore, due to the physical inability to measure movement endpoints on the keys, the computation of the effective width is troublesome. As such, an alternative approach was proposed by Soukoreff and MacKenzie (2004) based on error rate, as presented in (Equation 5) and used in this research. Here, Err is the error rate of a specific condition that equals the number of wrongly pressed keys over the total number of pressed keys, and $z(x)$ represents "the inverse of the standard normal cumulative distribution, or, the z-score that corresponds to the point where the area under the normal curve is x%." These accuracy adjustments must be performed for each individual condition

and participant, given that W_e describes the "within-participant variability," and hence pooling endpoint information will not result in proper results (Soukoreff & MacKenzie, 2004).

$$ID_e = \log_2\left(\frac{A_{ij}}{W_{ij_e}} + 1\right)$$

where

$$W_{ij_e} = \begin{cases} \min\left(H_j, W_j\right) \times \dfrac{2.006}{z(1 - Err/2)} & \text{if } Err > 0.0049\% \\ \min\left(H_j, W_j\right) \times 0.5089 & \text{otherwise} \end{cases}. \quad (5)$$

Touch-Based Navigation Display (TND)

Research by Bi, Li, and Zhai (2013) has extended the original Fitts' law to produce the Finger Fitts' Law, shown in Equation 6. Their research proved effective in modeling finger input using touchscreens. Two new parameters are introduced: σ, the variation in movement endpoints, and σ_a, the variation in input device precision (e.g., finger width). The former is calculated using the distribution in endpoint coordinates during the task, where a bivariate standard deviation σ_{xy} is used for two-dimensional (2D) movements. The latter can be measured using a finger calibration task, where users are asked to repeatedly touch an identical (in size, not location) target; exact touch locations are used in this research to calculate the bivariate standard deviation σ_{xy} instead of σ.

$$MT = a + b\log_2\left(\frac{A}{W_e} + 1\right) \quad \text{where } W_e = \sqrt{2\pi e\left(\sigma_{xy}^2 - \sigma_a^2\right)} \quad (6)$$

METHOD

The objective of the experiment was to develop and compare Fitts' law models for each of the three interfaces using the respective models described earlier. The experiment consisted of three separate, but similar sub-experiments corresponding to the interfaces. The overarching design of the experiment is discussed here, followed by a brief discussion of each sub-experiment focusing on one interface. Each experiment explicitly measured the effect of ID on the observed MT for participants engaged in an aimed rapid movement task using the respective interface.

Participants

Given that the goal of the experiment was to describe human performance in performing a precision pointing task for a specific interface using Fitts' law, prior experience with piloting aircraft and/or interacting with the interfaces was not relevant. The lack of previous encounters with either the MCP or CDU (for example by naive participants) was dealt with during a training phase, where each participant got sufficiently accustomed to the input device (see "General Procedure"). Right-handed participants were preferred given the positioning in the left seat and thus interface operation with the right hand. A total of 14 people participated in the experiment, of which a brief profile is given in Table 1. Note that one left-handed participant was invited in order to see the effect of handedness in using the TND.

Experiment Design

The experiment had a within-participants design. Figure 2 illustrates the different orders employed in presenting the conditions for 12 participants. Three groups of four participants (A, B, C) were administered the same interface order. The remaining two participants followed the order of the first two groups and of which one was left-handed. Given that each interface was different, different manipulations were required to achieve comparable indices of difficulty. The design was such that number of repetitions per unique ID per participant ranged between 10 and 12, similar to that found and recommended in literature (Accot & Zhai, 1997; Bi et al., 2013; Soukoreff & MacKenzie, 2004; Stoelen & Akin, 2010). The specific manipulations per interface condition will be detailed in the description of the interface conditions. Finally, the ranges of the evaluated inputs per interface condition were representative for a realistic LNAV re-routing task to avoid a

TABLE 1: Profile of Participants

Profile	13 students, 1 professor
Gender	11 male, 3 female
Age	Ranging 21 to 49, averaging 24 years
Handedness	13 right-handed, 1 left-handed

weather cell (i.e., dialing in a heading with the MCP, inserting a new waypoint using the CDU, and finger dragging a waypoint using the TND).

Apparatus

The experiment was conducted in the SIMONA Research Simulator (SRS) at the Delft University of Technology, shown in Figure 3. Motion and outside visual capabilities were not utilized; however, the interior cabin provided a realistic look and feel to the interaction between participants and the three flight deck interfaces. Similar to a real flight deck, the locations and sizes of the interfaces, as well as the position of the participants in the left seat, and left of the interfaces, were fixed.

Due to space confinements in the SIMONA Research Simulator, the touchscreen was located below the CDU (see Figure 3). As a result, to allow for a proper comparison, the participants were required to put their seat backwards when using the TND. Markers were installed on the cabin floor to ensure constant seat positioning. As such, the participant's relative location to the touchscreen was comparable to that of the CDU and MCP.

General Procedure

This research complied with the tenets of the Declaration of Helsinki and was approved by the Human Research Ethics Committee of TU Delft. Informed consent was obtained from each participant. Participants received a briefing document a few days prior to the experiment. An introduction was given concerning the relevance of the experiment, the task to be conducted, and the expected time schedule. Prior to each interface condition, following a standardized procedure, a verbal briefing was given. Most importantly, and "essential for any Fitts' law experiment" (Soukoreff & MacKenzie, 2004), the participant was requested to put specific emphasis on *speed and accuracy* in order to achieve an approximate 96% target hit-percentage with a smooth consistent input motion. Feedback on actual hit-rates was provided during all runs of the experiment. Training runs preceded data measurement and provided participants with time to master the speed-accuracy trade-off. When they reached the 96% target, they were considered to be sufficiently trained. More details on specific procedures per interface condition will be provided later.

Mode Control Panel (MCP) Condition

The MCP setup is presented in Figure 4. On the inboard screen, the navigation display (❶) is shown, on which the task information was presented. A magenta heading bug (see ❷) indicates the heading commanded on the MCP. At the start of each trial, the bug was reset to the north-up position. Two independent variables were used: the angular amplitude α and angular width ω; together they determine the ID. The target was shown using two cyan lines (see ❸), the angular distance between which represents the width ω. The angular distance between the starting position of the heading bug and the center of the target is the movement amplitude α.

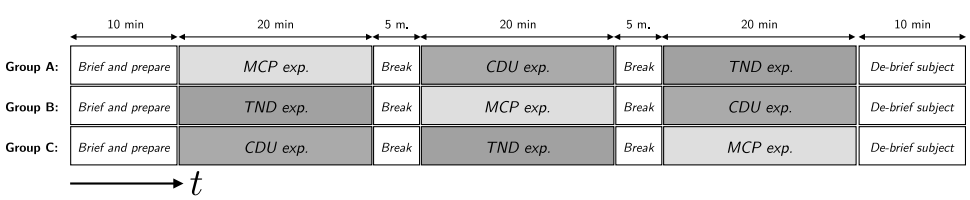


Figure 2. Schedule of the experiment per participant group.
Note. MCP = mode control panel; CDU = control display unit; TND = touch-based navigation display.

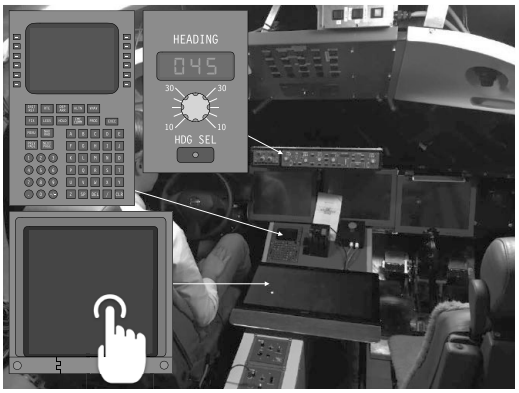


Figure 3. Cabin of the SIMONA Research Simulator (SRS) showing each of the three flight deck interfaces.

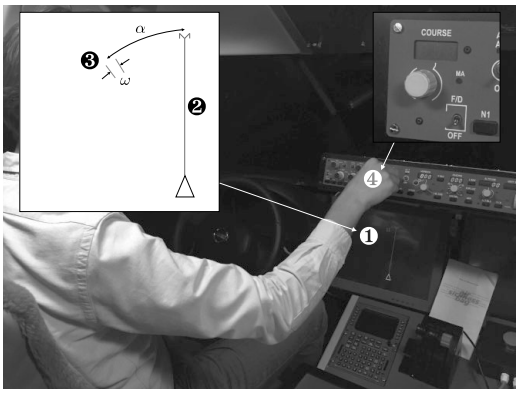


Figure 4. Experiment procedure for mode control panel (MCP) experiment, showing an illustration of the navigation display and heading control knob.

The choice of α and ω were such that they form a representative and realistic range of IDs for the MCP (e.g., when circumnavigating complex weather systems):

$$\alpha = [10, 20, 30, 40, 50, 60][\text{deg}]$$
$$\omega = [2, 4, 6, 8][\text{deg}]$$

These combinations resulted in an ID range of $[1.17, 4.95]$. In total, participants were confronted with 24 (6×4) different combinations. Given that the variables α and ω are multiples of two, a total of 16 unique ID values existed. For example, the combinations $\alpha = 10, \omega = 2$ and $\alpha = 20, \omega = 4$ produce the same ID.

Participants needed to use the course select rotary knob to hit the target ω at a certain amplitude α. The course select knob (illustrated by ❹) on the MCP is a standard rotary encoder with 24 "clicks" per full rotation. Note that for this study, the course knob was used due to a malfunction in the heading knob. Because both knobs operate in the same fashion (although they are used in a different navigation context) and initial hand movements toward the knob was not included in movement time measurements, this was not considered problematic. A small LCD display above the knob reflected the commanded heading. The movement time MT to hit the target was measured in milliseconds. In accordance with recommendations in literature (Soukoreff & MacKenzie, 2004) only the actual time the participant moved the heading knob was measured, thereby omitting engage, homing, dwell, and reaction times. Hereby, confounding factors such as cognitive effort required to understand the task and initial hand movements toward the interface were mitigated. Accuracy was measured by recording the physical endpoints of each individual movement. During the experiment the success rate in acquiring the target was displayed in the control room and communicated to the participant to provide feedback on their adherence to the speed-accuracy trade-off governing Fitts' law.

The training phase for the MCP condition contained one full set of 24 combinations. The measurement phase constituted eight sets of 24 combinations, totaling 192 measurements runs.

CDU Condition

The experiment setup is shown in Figure 5. An illustration of the CDU including the display is shown in ❶. In a re-routing LNAV task, pilots use the CDU to insert new waypoints by entering their name in the scratchpad (see ❷) and inserting it in the list of waypoints through one of the line select keys (LSK; see ❸). The CLR key could be used to backspace the scratchpad. The full content of the scratchpad could be inserted to any of the 12 line select keys (see ❸) by pushing the respective key. The text subsequently moved and the scratchpad was cleared.

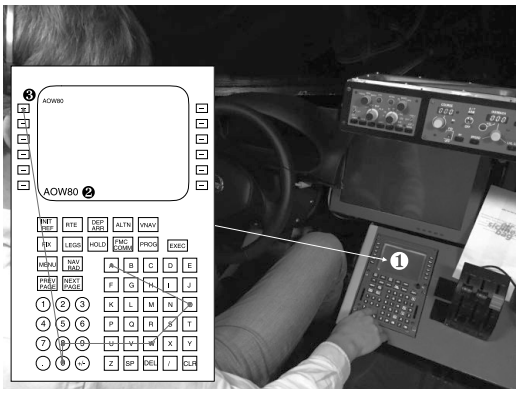


Figure 5. Experimental setup for control display unit (CDU) experiment, showing an illustration of the control display unit and location within the flight deck.

In this experiment, the variables A and W were defined by a set of words that needed to be entered and subsequently moved to target line select keys. Figure 5 shows an example where a participant is required to enter the word AOW80 before moving it to the top-left line select key. The amplitude A was characterized as the shortest distance between each key, and the width W was characterized as the minimum of either the height or width of the key (MacKenzie & Buxton, 1992). The participant was instructed to first search and find the necessary keys prior to initiating data entry in order to reduce cognitive effort. This meant that one five letter word that needed to be positioned at a specified line select key constituted five movements with respective A and W values, because the time needed to search for the first key was not recorded.

In order to complete the model, one word consisted of repeated keys in order to determine MT_{repeat} as shown in Equation 4. The set of words and LSKs were carefully chosen to encompass a representative range of indices of difficulty for a realistic LNAV task.

Words=[KLM19, AET50, 47MAY
SSSSS, DJS73, ANW80][-]
Target LSK=[L1,L2,L6,R2,R4,R5][-],

An accurate technical drawing of the CDU used during the experiment was consulted to calculate A and W and resulted in an ID range of $[1.26, 5.17]$.

The combinations of words and LSKs provided a total set of 36 different conditions, each of which consisted of five Fitts' law movements. Therefore, a minimum of 180 Fitts' law measurements could be made during one set of combinations.

Similar to the MCP experiment, the movement time MT, excluding homing, dwell, and reaction times, was measured in milliseconds. The accuracy, measured as the number of correct inputs divided by the total amount of keystrokes, was measured and used to provide as feedback to participants. Endpoint distributions of the inputs (i.e., finger locations) on the keys could not physically be measured, however. The training consisted of one block of all 36 conditions (in a random order) and the measurement phase featured two blocks of 36 conditions.

TND Condition

The experiment setup is shown in Figure 6. A large touchscreen was installed horizontally on the center pedestal of the SRS cockpit. An illustration of the display presented on the screen is shown in ❶. A white object was shown with a magenta crosshair at its center (see ❷), which could be moved around using touch-based input.

The target was depicted using a cyan circle (see ❸) with a black crosshair. The distance to be traversed, the amplitude A, and the diameter or width W of the circular target constituted the two variables that were manipulated. A representative and wide variety of variables A and W were selected. Finally, given that literature has found direction to be a confounding factor (Soukoreff & MacKenzie, 2004), a direction "heading" variable ϕ and display rotation variable θ were introduced, as illustrated in Figure 6. The rotation angle θ rotates the entire reference frame of the display.

$$A = [45, 80, 115, 150][\text{mm}]$$
$$W = [5, 15, 25, 35][\text{mm}]$$
$$\phi = [-25, 0, 25][\text{deg}]$$
$$\theta = [0, 90, 180, 360][\text{deg}]$$

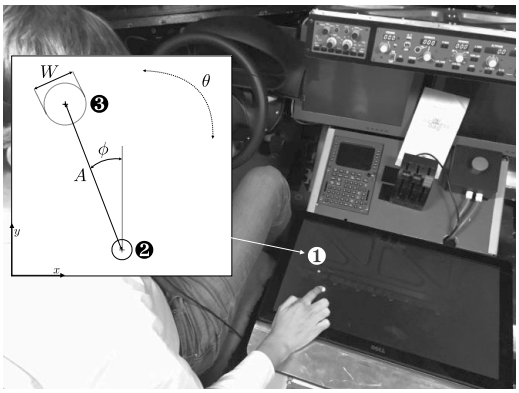


Figure 6. Experimental apparatus for the touch-based navigation display (TND) experiment, showing an illustration of the touchscreen display and its location on the flight deck.

The choice in variables resulted in a total set of 192 different input combinations. However, only 16 (4×4) different combinations of A and W were present due to the use of directional variables. As a result, the ID range was $[1.19, 4.95]$ and thus comparable to the MCP and CDU indices of difficulty. Similar to the MCP and CDU conditions, the TND input combinations were designed to represent inputs that can be expected during a realistic LNAV rerouting task using a touchscreen device.

Consistent with the MCP and CDU experiments, the MT, excluding homing, dwell, and reaction times, was measured in milliseconds. Accuracy was the other dependent variable and was measured by recording physical endpoints of each individual movement. Given the 2D nature of the task, a bivariate endpoint standard deviation (σ_{xy}) was used, which has been found to better describe 2D Fitts' law tasks (Wobbrock, Cutrell, Harada, & MacKenzie, 2008).

For the finger calibration task to calibrate σ_{xy}, a fixed diameter magenta target, slightly larger than a typical index finger, with a white crosshair was drawn at a random (x, y) location on the display.

The task setup in both the training and measurement phases were equal. Once the participant was ready, a set of 192 conditions were loaded, and both the object and target were reset to their respective positions. Measurement started when the participant had successfully acquired the object and started to move it. Object acquisition was done by providing a touch input within a touch area equal in size and location of the object. During the experiment, the success rate in acquiring the target was displayed in the control room and communicated to the participant to provide valuable feedback on their adherence to the speed-accuracy trade-off governing Fitts' law.

Training consisted of 192 runs containing all possible combinations, and the measurement phase features again 192 runs, albeit in a different (randomized) order.

RESULTS

The numerical results of the three interface conditions are summarized in Table 2, the model fits are shown in Figure 7, and the distributions of accuracy values per interface are depicted in Figure 8. For the MCP condition, the proposed adjustment based on accuracy was done by computing the effective width W_e based on the actual distribution of movement endpoints per ID. Based on the effective width W_e, an effective index of difficulty ID_e was calculated (circles in Figure 7). An analysis of variance (ANOVA) test revealed a significant effect of ID on MT, $F(4.33, 56.28) = 138.47, p < .01$. Compared with the other interfaces, the MCP has the lowest y-intercept and the lowest throughput. In terms of accuracy, an ANOVA reported a significant effect of the interface conditions $F(1.134, 14.74) = 5.623, p < .05$. Pairwise comparisons (adopting a Bonferroni correction) only reported a significant difference between the MCP and CDU. From Figure 8, it can be observed that the MCP accuracy scored between the CDU and the TND.

For the CDU condition, the proposed adjustment for accuracy was done by computing the effective width W_e based on the error percentages per ID. Based on this effective width W_e, an effective index of difficulty ID_e was calculated (plusses in Figure 7). An ANOVA test showed a significant effect of ID on mean MT, $F(3.76, 48.83) = 37.05$, $p < .01$. The CDU had the highest throughput as well as the highest accuracy. In Figure 8, it can also be seen that the variability in achieved accuracy is smallest compared with the other interfaces.

TABLE 2: The Fitts' Law Model, Quality of Fit (R^2), Mean Accuracy and Throughput of Each Interface

Interface	Fitts' Law Model (ms)	R^2	Accuracy (%)	Throughput (Bits/s)
MCP	$MT = 154.8 + 494.7 \cdot ID_e$	0.969	96	1.80
CDU	$MT = 337.9 + 91.7 \cdot ID_e$ $MT_{repeat} = 267.9$	0.835	99	5.20
TND	$MT = 212.3 + 180.3 \cdot ID_e$	0.879	95	3.88

Note. MCP = mode control panel; CDU = control display unit; TND = touch-based navigation display.

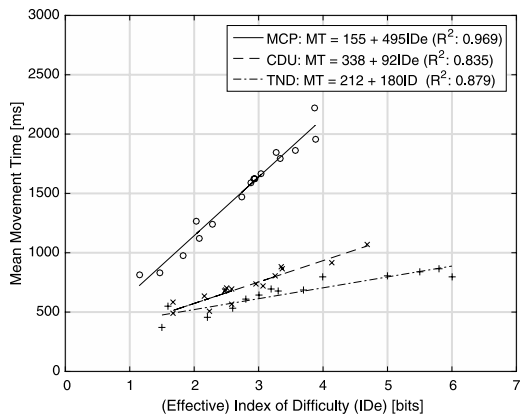


Figure 7. Final Fitts' law models of each individual interface plotted on the same graph for comparative purposes.
Note. MCP = mode control panel; CDU = control display unit; TND = touch-based navigation display.

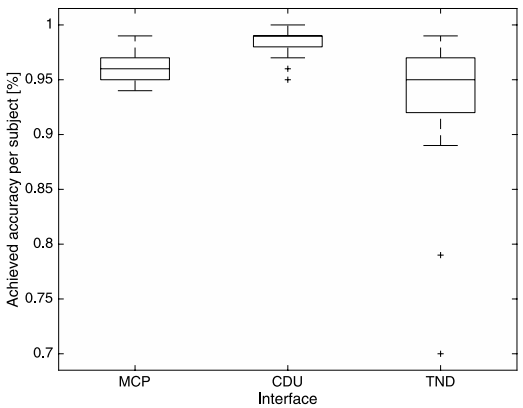


Figure 8. Observed accuracy scores per participant per experiment.
Note. MCP = mode control panel; CDU = control display unit; TND = touch-based navigation display.

For the TND condition, the proposed adjustment for accuracy was done by computing the effective width W_e as shown in Equation 6. Based on W_e, an effective index of difficulty ID_e was calculated (crosses in Figure 7). An ANOVA test concluded that there was a significant effect of ID on MT, $F(2.67, 34.67) = 37.20, p < .01$. From the dedicated calibration tests, the finger calibration parameter σ_a was 3.6 mm. According to Figure 8, the TND has the lowest average accuracy as well as the highest spread pattern in accuracy. Although these results were not found to be significant compared to the MCP and CDU, this does show that navigation-type inputs with the TND can be more error prone compared to the conventional flight deck interfaces. In throughput, however, the TND scores better than the MCP.

DISCUSSION

The results of all three interface conditions show that the different variations of Fitts' law, acquired from literature and introduced in this article, are adequate ways to develop and compare accuracy and throughput models for the MCP, CDU, and a TND. This is illustrated by Figure 7, and the R^2 fit qualities of 0.97, 0.84, and 0.88 for the aforementioned interfaces, respectively. The good fit for the MCP was a pleasant surprise, given that literature lacks a study looking at the applicability of Fitts' law to a rotary controller providing *discrete* input signals.

Furthermore, when scrutinizing the y-intercept parameter (a) of each Fitts' law model, the CDU indeed results in the largest expected movement time for tasks of zero difficulty, namely 338 ms. The TND follows with 212 ms.

Interestingly, the MT_{repeat} of the CDU is different from its y-intercept, despite that both represent zero index of difficulty ($ID = 0$). This is, however, consistent with the findings of Soukoreff and MacKenzie (2002), who indicated that the y-intercept for keyboard-entry tasks is a "theoretical" movement time based on linearly regressing data containing inter-key movements ($ID > 0$ bits) and may therefore not accurately describe realistic movement times for pressing the same key twice. During the CDU trials, it was also observed that a significant amount of force was required to successfully press the keys. Participants were even observed to occasionally continue toward a next key after unsuccessfully hitting the previous one. Furthermore, although participants were requested to search the necessary keys before initiating data entry, the cognitive effort required to find the required keys is expected to still affect the movement time. The lower y-intercepts for the TND and MCP do suggest that they are indeed more direct (and perhaps more intuitive) in their use.

In terms of accuracy, the MCP scores similar to the TND, but the variability is much smaller than for the heading control knob on the MCP. This finding is very similar to that of Stanton et al. (2013), who also compared a rotary controller with a touchscreen. Interesting to note is that even though data were only available from one left-handed participant, these scores were 70% on the TND compared to 96% and 99% on the MCP and CDU, respectively. This suggests that using the traditional interfaces with a nondominant hand is easier than with a touchscreen. This finding is intriguing, given that the pilot position within the flight deck relative to interfaces is fixed and cannot easily be adjusted by the pilot. Further research on the effect of handedness on flight deck performance is therefore warranted. Following discussions with participants and observations made during the experiment, the accuracy results could also have been attributed to the tactile nature of and the fixed physical locations of the dials and buttons on the traditional interfaces. Due to the lack of tactile feedback and high freedom of movement with the touchscreen, precise inputs were sometimes more difficult to achieve.

Regarding throughput, scores were highest with CDU, followed by the TND and MCP, respectively. According to Figure 7, the TND and CDU result in similar movement times at low ID values (i.e., 1.5) as their Fitts' law models converge. At higher ID values, however, the lines diverge and the TND is at a disadvantage compared to the CDU. Based on these results it can be said that for a given short time interval, the CDU can handle more difficult tasks compared to the other two interfaces. This may be explained by the calculation of ID, which is defined by the movement amplitude and target width. On the CDU the target width remained constant, given that the keys had a pre-defined size. Hence, the difficulty in movements was reflected in the distance to be moved. Thus, moving a larger distance was observed to be easier than acquiring a very narrow target, which is reflected by Equation 4. In addition, the physical keys on the CDU make it fairly easy to acquire the target successfully. On the contrary, with the MCP and TND, target difficulty varied both by amplitude and width. For the latter, it was observed on both interfaces that a very narrow target slowed down participants and required them to be more accurate. Finally, movement times were found to be substantially longer for the MCP than for the other two interfaces. This may be attributed to the latency and nonlinear movement of the heading control knob noted by several participants. Research by Stanton et al. (2013) also found that use of a rotary controller resulted in longer task times compared to a touchscreen interface.

Although scores with the CDU were highest on both accuracy and throughput, this does not imply that it is therefore the most optimal interface with the FMS. During the experiment, participants were asked to locate the necessary keys prior to key entry to keep cognitive effort at a minimum. Hence, good performance with the CDU reflects that the user is fully aware of the necessary steps to execute. However, during a more complex task, a substantial amount of cognitive effort is expected in determining the necessary actions with the CDU. Thus, the question remains whether, when used during a more realistic navigation task, the CDU is still better than a touch-based interface.

In addition, during a realistic LNAV task, for example, to avoid bad weather, pilots generally use both the CDU and MCP. In most cases, however, pilots will not use these interfaces concurrently. That is, they use the MCP to deviate from the planned route by dialing in a heading to fly around a weather cell and finally use the CDU to fly directly toward the nearest route waypoint when they cleared the weather cell. On one hand, it can be said that our results could shed light on the expected total task difficulty and completion times for realistic flight navigation tasks requiring combined inputs, given the current focus on modeling the accuracy and throughput of the interfaces in isolation. On the other hand, our results may not be as simple as summing the throughput values and task completion times. In combined inputs with multiple interfaces, time delays associated with redirecting hand movements, distributing visual attention over multiple interfaces, time to engage, homing, and dwell will also play important roles. How such combined interactions with two different interfaces at separate locations on the flight deck compare to a TND, and to what extent our obtained Fitts' models can predict the results of such interactions, is therefore worth exploring further in a follow-on experiment.

KEY POINTS

- The accuracy and throughput characteristics of three flight deck interfaces, that is, the MCP, the CDU, and a TND, were accurately modeled with Fitts' law.
- The Fitts' law analysis showed the CDU as most effective in both accuracy and throughput, which indicates that more difficult tasks can be handled better with the CDU within a short time frame.
- Although the Fitts' law models derived in this research described individual input movements, they may enable improved analysis and prediction of total task difficulty and completion times for realistic flight navigation tasks that would require a series of combined movements.

ORCID iD

Daan M. Pool https://orcid.org/0000-0001-9535-2639

REFERENCES

Accot, J., & Zhai, S. (1997). Beyond Fitts' law: Models for trajectory-based HCI tasks. *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 295–302). Atlanta, GA. doi:10.1145/258549.258760

Ballas, J. A., Heitmeyer, C. L., & Pérez-Quiñones, M. A. (1992). Evaluating two aspects of direct manipulation in advanced cockpits. *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 127–134). Monterey, CA. doi:10.1145/142750.142770

Bi, X., Li, Y., & Zhai, S. (2013). FFitts law: Modeling finger touch with Fitts' law. *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 1363–1372). Paris, France. doi:10.1145/2470654.2466180

Bjørneseth, F. B., Dunlop, M. D., & Hornecker, E. (2012). Assessing the effectiveness of direct gesture interaction for a safety critical maritime application. *International Journal of Human-Computer Studies*, *70*, 729–745. doi:10.1016/j.ijhcs.2012.06.001

Bulfer, B. (1991). *Big Boeing FMC user's guide*. Kingwood, TX: Leading Edge.

Degani, A., Palmer, E., & Bauersfeld, K. G. (1992). "Soft" controls for hard displays: Still a challenge. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *36*, 52–56. doi:10.1177/154193129203600114

Dodd, S. R., Lancaster, J., Grothe, S., DeMers, B., Rogers, B., & Miranda, A. (2014). Touch on the flight deck: The impact of display location, size, touch technology & turbulence on pilot performance. *Proceedings of the IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC)*. doi:10.1109/DASC.2014.6979428

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*, 381–391. doi:10.1037/h0055392

Flach, J. M., Hagen, B. A., O'Brien, D., & Olson, W. A. (1990). Alternative displays for discrete movement control. *Human Factors*, *32*, 685–695. doi:10.1177/001872089003200606

Forlines, C., Wigdor, D., Shen, C., & Balakrishnan, R. (2007). Direct-touch vs. mouse input for tabletop displays. *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 647–656). San Jose, CA. doi:10.1145/1240624.1240726

Francis, G., & Oxtoby, C. (2006). Building and testing optimized keyboards for specific text entry. *Human Factors*, *48*, 279–287. doi:10.1518/001872006777724390

Gao, Q., & Sun, Q. (2015). Examining the usability of touch screen gestures for older and younger adults. *Human Factors*, *57*, 835–863. doi:10.1177/0018720815581293

Huisman, H., Verhoeven, R. P. M., Van Houten, Y. A., & Flohr, E. L. (1997). Crew interfaces for future ATM. *Proceedings of the AIAA/IEEE digital avionics systems conference*. Irvine, CA. doi:10.1109/DASC.1997.636189

Hutchins, E. L., Hollan, J. D., & Norman, D. A. (1985). Direct manipulation interfaces. *Human-Computer Interaction*, *1*, 311–338. doi:10.1207/s15327051hci0104_2

Jagacinski, R. J., & Fisch, J. (1997). Information theory and Fitts' law. In J. M. Flach & R. J. Jagacinski (Eds.), *Control theory for humans* (pp. 17–26). Mahwah, NJ: Lawrence Erlbaum.

Jagacinski, R. J., Repperger, D. W., Ward, S. L., & Moran, M. S. (1980). A test of Fitts' law with moving targets. *Human Factors*, *22*, 225–233. doi:10.1177/001872088002200211

Jax, S. A., Rosenbaum, D. A., Vaughan, J., & Meulenbroek, R. G. J. (2003). Computational motor control and human factors: Modeling movements in real and possible environments. *Human Factors*, *45*, 5–27. doi:10.1518/hfes.45.1.5.27226

Kaminani, S. (2011). Human computer interaction issues with touch screen interfaces in the flight deck. *Proceedings of the AIAA/IEEE digital avionics systems conference*. Seattle, WA. doi:10.1109/DASC.2011.6096098

MacKenzie, S. I. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction*, *7*, 91–139. doi:10.1207/s15327051hci0701_3

MacKenzie, S. I., & Buxton, W. (1992). Extending Fitts' law to two-dimensional tasks. *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 219–226). Monterey, CA. doi:10.1145/142750.142794

Rogers, W. A., Fisk, A. D., McLaughlin, A. C., & Pak, R. (2005). Touch a screen or turn a knob: Choosing the best device for the job. *Human Factors*, *47*, 271–288. doi:10.1518/0018720054679452

SESAR Joint Undertaking. (2014). *i4D and SESAR*. Retrieved from https://www.sesarju.eu/taxonomy/term/66

Shneiderman, B. (1982). The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology*, *1*, 237–256. doi:10.1080/01449298208914450

Soukoreff, R. W., & MacKenzie, S. I. (1995). Theoretical upper and lower bounds on typing speed using a stylus and a soft keyboard. *Behaviour & Information Technology*, *14*, 370–379. doi:10.1080/01449299508914656

Soukoreff, R. W., & MacKenzie, S. I. (2002, May). *Using Fitts' law to model key repeat time in text entry models*. Poster presented at Graphics Interface—GI2002. Retrieved from http://soukoreff.com/academic/GI02-Poster.pdf

Soukoreff, R. W., & MacKenzie, S. I. (2004). Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *International Journal of Human-Computer Studies*, *61*, 751–789. doi:10.1016/j.ijhcs.2004.09.001

Stanton, N. A., Harvey, C., Plant, K. L., & Bolton, L. (2013). To twist, roll, stroke or poke? A study of input devices for menu navigation in the cockpit. *Ergonomics*, *56*, 590–611. doi:10.1080/00140139.2012.751458

Stoelen, M. F., & Akin, D. L. (2010). Assessment of Fitts' law for quantifying combined rotational and translational movements. *Human Factors*, *52*, 63–77. doi:10.1177/0018720810366560

Stuyven, G., Damveld, H. J., & Borst, C. (2012). Concept for an avionics multi touch flight deck. *SAE International Journal of Aerospace*, *5*, 164–171. doi:10.4271/2012-01-2120

Trudeau, M. B., Udtamadilok, T., Karlson, A. K., & Dennerlein, J. T. (2012). Thumb motor performance varies by movement orientation, direction, and device size during single-handed mobile phone use. *Human Factors*, *54*, 52–59. doi:10.1177/0018720811423660

van Marwijk, B. J. A., Borst, C., Mulder, M., Mulder, M., & van Paassen, M. M. (2011). Supporting 4D trajectory revisions on the flight deck: Design of a human–machine interface. *The International Journal of Aviation Psychology*, *21*, 35–61. doi:10.1080/10508414.2011.537559

Wobbrock, J. O., Cutrell, E., Harada, S., & MacKenzie, S. I. (2008). An error model for pointing based on Fitts' law. *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 1613–1622). Florence, Italy. doi:10.1145/1357054.1357306

Nout C. M. van Zon is a consultant at Simon-Kucher & Partners. He received his MSc degree (cum laude) in aerospace engineering in 2017 from Delft University of Technology.

Clark Borst is an assistant professor at Delft University of Technology, Faculty of Aerospace Engineering, Department of Control & Operations. He received his PhD degree in aerospace engineering in 2009 from Delft University of Technology.

Daan M. Pool is an assistant professor at Delft University of Technology, Faculty of Aerospace Engineering, Department of Control & Operations. He received his PhD degree (cum laude) in aerospace engineering in 2012 from Delft University of Technology.

Marinus M. van Paassen is an associate professor at Delft University of Technology, Faculty of Aerospace Engineering, Department of Control & Operations. He received his PhD degree in aerospace engineering in 1994 from Delft University of Technology.

Date received: October 1, 2018
Date accepted: June 12, 2019

Effects of School Backpacks on Spine Biomechanics During Daily Activities: A Narrative Review of Literature

Cazmon Suri, Iman Shojaei, Babak Bazrgari,
University of Kentucky, Lexington, USA

Objective: The purpose of this narrative review is to summarize the effects of carrying school backpacks on spine and low-back biomechanics as a risk factor for low back pain in young individuals.

Background: Backpacks constitute a considerable daily load for schoolchildren. Consistently, a large number of children attribute their low back pain experience to backpack use.

Method: A literature search was conducted using a combination of keywords related to the impact of carrying backpacks on lower back biomechanics. The references of each identified study were further investigated to identify additional studies.

Results: Twenty-two studies met inclusion criteria. A total of 1,159 people aged 7 to 27 years were included in the studies. The added load of a backpack and the changes in spinal posture when carrying a backpack impose considerable demand on internal tissues and likely result in considerable spinal loads. The findings included results related to the effects of backpack weight and position on trunk kinematics and spine posture as well as trunk muscle activity during upright standing, walking, and ascending and descending stairs.

Conclusion: Backpack-induced changes in trunk kinematics for a given activity reflect alterations in mechanical demand of the activity on the lower back that should be balanced internally by the active and passive responses of lower back tissues. Although the reported alterations in trunk muscle activities and lumbar posture are indications of changes in the active and passive response of the lower back tissues, the resultant effects on spinal load, that is, an important causal factor for low back pain, remains to be investigated in the future. A knowledge of backpack-induced changes in spinal loads can inform design of interventions aimed at reduction of spinal load via improved backpack design or limitation on carrying duration.

Application: This narrative review is intended to serve as an educational article for students and trainees in ergonomics and occupational biomechanics.

Keywords: narrative review, kinematics, children, low back pain, posture, backpack

Address correspondence to Babak Bazrgari, F. Joseph Halcomb III, M.D. Department of Biomedical Engineering, University of Kentucky, 514E Robotic and Manufacturing Building, Lexington, KY 40506, USA; e-mail: babak.bazrgari@uky.edu.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 909–918
DOI: 10.1177/0018720819858792

INTRODUCTION

Low back pain (LBP) is a growing concern for young people with 40% of 9- to 18-year-olds across the world reporting they have had LBP (Calvo-Munoz, Gomez-Conesa, & Sanchez-Meca, 2013; Negrini & Carabalona, 2002). Reported annual prevalence of LBP ranges from 22% to 51% in children aged 12 to 16 years (Watson et al., 2002) and is responsible for missed school days and sleeping problems in 20% and 50% of children, respectively (Roth-Isigkeit, Thyen, Stoven, Schwarzenberger, & Schmucker, 2005).

The weight carried in a backpack has been suggested to play a pathogenic role in the development of LBP in children (Negrini & Carabalona, 2002; Nicolet, Mannion, Heini, Cedraschi, & Balague, 2014). Furthermore, 82% of children aged 11 to 14 years with LBP attribute their pain to backpack use (Shymon et al., 2014). Backpack loads for young people have increased over the past two decades, raising concerns among medical practitioners and parents about the possible detrimental effects (Al-Khabbaz, Shimada, & Hasegawa, 2008) on their health. Recent studies from different countries have shown average backpacks in school children are heavier than the recommended amount of 10% to 22% body weight (BW) (Brzek et al., 2017; Mackenzie, Sampath, Kruse, & Sheir-Neiss, 2003; Negrini & Carabalona, 2002; Sheir-Neiss, Kruse, Rahman, Jacobson, & Pelli, 2003; Whittfield, Legg, & Hedderley, 2001). This is concerning because LBP at a young age has been suggested to play an important role in developing chronic LBP in adulthood (Negrini & Carabalona, 2002). Despite concerns regarding the negative health effects of heavy backpacks, there is limited knowledge about the mechanism(s) linking carrying a heavy backpack

with development of LBP in young people to inform suitable interventions.

Repetitive loading of the lumbar spine increases the risk of LBP through cumulative or overuse injuries to spinal tissues (Adams, 2013; Mackie & Legg, 2008). Both the frequency and magnitude of loads acting on the spine contribute to risk of cumulative injuries (Brinckmann, Biggemann, & Hilweg, 1988). Backpacks constitute a considerable daily "occupational" load for schoolchildren (Shymon et al., 2014); backpacks are often carried during repetitive or prolonged activities of daily living such as standing, walking, jogging, and stair climbing. Under such conditions, spinal loads are likely to increase considerably. The added mechanical demands of the backpack load on the lower back alters trunk muscle response and recruitment (e.g., involving coactivation) because of muscle fatigue and/or spinal instability (Cholewicki, Panjabi, & Khachatryan, 1997; Granata & Orishimo, 2001; Potvin & O'Brien, 1998).

Vertebrae ossification is not complete until the mid-20s and the relatively high amount of cartilage in the skeletons of children put them at greater risk of overuse injuries compared with adults (O'Day, 2008). Vulnerability of cartilage to shear stresses and repetitive trauma decrease soft-tissue flexibility and induce muscle imbalances (O'Day, 2008). Given that overuse injuries in spinal tissues are likely to have a role in developing LBP, the objective of this review is to summarize the findings of studies that have investigated the effects of carrying school backpacks on the lower back mechanics of young people. Specifically, how backpacks affect forces and deformations of lower back tissues—also referred to as mechanical environment of the lower back throughout this review—will be examined.

METHOD

A literature search was conducted to identify all pertinent research studies regarding the effects of backpacks on the spine and lower back biomechanics among young people. InfoKat Discovery, a search engine offered by the University of Kentucky library system, was used to search for peer-reviewed articles using combinations of keywords (Table 1) in the title or abstract. InfoKat Discover searches many scientific and medical databases including PubMed, CINAHL, and Ei Compendex. Initial screening identified studies with at least one keyword from each category in its title or abstract. The reference lists from identified articles were checked for additional sources. The first author conducted the search and provided initial screening of the identified literature. Assistance from the coauthors was provided in secondary screening of articles to assure relevance to this review. Specifically, the inclusion criteria considered for this review were (a) reporting of biomechanical measures related to the lower back and (b) involving backpack use during activities of daily living. Some studies had outcome measures in addition to those related to the lumbar region of the back and were also included in this review. Due to the small number of papers meeting these criteria, no limit on the publication year was set. Information regarding sample size, age, gender, backpack type, loading type and location, load (%BW), measurement method, task and duration, and outcome measures were extracted from the final set of articles and are summarized in Table 2.

All articles were further screened to exclude any studies that did not investigate biomechanical measures in young people related to carrying a backpack. Because most of the identified backpack studies focused on the effects of weight or position of the backpack on the lower back and spine biomechanics during upright standing, walking, and ascending and descending stairs, this review has been organized to present relevant findings according to these variables.

RESULTS

The initial database search identified 42 papers, of which 22 met our criteria and were included in the review (Table 2).

Effects of Backpack Weight

Posture and kinematics. Backpack loading has been reported to affect deformation of lumbar disks (i.e., L1-L2, L4-L5, and L5-S1) with a positive association between loading and deformation (Shymon et al., 2014). Specifically, a backpack load of ~4 kg, relative to no-load condition, was found to cause ~ 13% decrease in the height of anterior region of the L5-S1

TABLE 1: Four Groups of Keywords Were Used to Search the Databases

Keyword Group 1	Keyword Group 2	Keyword Group 3	Keyword Group 4
Backpack	Stress	Back	Children
Schoolbag	Strain	Lower Back	Teen/Teenager
Book Bag	Shear	Trunk	Young
Book Pack	Kinematics	Lumbar	Adolescences
Demand	Kinetics	Pelvis	School
Carriage	Biomechanics	Spine/Spinal	Juvenile
—	Posture	—	—

Note. During the initial search, an article would qualify for additional screening if its title or abstract contained at least one keyword from each category.

intervertebral disk. Backpack-induced alterations in lumbar posture have been suggested to adversely affect repositioning ability of the lumbar spine (Brzek et al., 2017; Chow, Leung, & Holmes, 2007; Pascoe, Pascoe, Wang, Shim, & Kim, 1997; Shymon et al., 2014).

In standing posture, backpack-induced forward trunk inclination, relative to a vertical line, has been reported to range between 3.02° and 6.8° for backpack weights ranging from 10% to 20% of BW (Brackley, Stevenson, & Selinger, 2009; Kistner, Fiebert, Roach, & Moore, 2013; Mackie & Legg, 2008). Backpack-induced forward trunk inclination has also been observed under lighter backpack weights (Ramprasad, Alias, & Raghuveer, 2010). Specifically, Ramprasad et al. (2010) studied 209 males of average age 12.5 years and reported an increase in forward trunk inclination of 3.21° compared with the no-backpack condition when using a backpack weight equal to 5% of BW (Ramprasad et al., 2010).

On the contrary, a study of 19 males with an average age of 21 years found an average trunk backward inclination of 3.43° during standing for backpack weights of up to 20% of BW (Al-Khabbaz et al., 2008). The conflicting results of this study may be due in part to the material used to increase the load. In the study by Al-Khabbaz et al. (2008), backpacks were filled with sand, which is more likely to collect at the bottom of the backpack, compared with weights or books such as in the study by Ramprasad et al. (2010).

In studies investigating the effects of backpack weight during standing, a negative relationship has been reported between lumbar lordosis (T12-L3-S1) and thoracic kyphosis (C7-T2-T5) and backpack weight. Specifically, an average of ~3° flattening in lumbar lordosis and thoracic kyphosis angles with 10% of BW increase in backpack weight has been shown (Chow et al., 2007; Walicka-Cuprys, Skalska-Izdebska, Rachwal, & Truszczynska, 2015). Negative relationships also were drawn between sacrum inclination and weight of backpack in 109 (58 girls and 51 boys) 7-year-old children such that an increase in backpack weight was associated with a decrease in sacrum inclination (backward pelvis tilt). The average difference in sacrum inclination between children used a backpack lighter than 10% of BW versus those who used a backpack heavier than 10% of BW was ~5° (Walicka-Cuprys et al., 2015).

During walking, trunk forward inclination has been reported to increase from 4.84° to 19.80° by increasing the backpack's weight from 10% to 20% of BW (Goodgold et al., 2002; Hong & Brueggemann, 2000; Hong & Cheung, 2003). Furthermore, backpack-induced forward inclination of the trunk during walking has been reported to increase not only by increases in backpack weight but also by increases in walking pace and distance (Goodgold et al., 2002; Hong & Brueggemann, 2000; Hong & Cheung, 2003). Li et al. investigated backpack-induced changes in trunk kinematics among 15 males with a mean age of 10.36 years and found that walking with a backpack heavier than 10% of BW induced a 4.55° increase in forward trunk inclination compared with the no-backpack

TABLE 2: Summary of the 22 Reviewed Studies Meeting Review Criteria, Sorted Alphabetically by Last Name of First Author

Study	Sample Size	Age (Years)	Gender	Backpack Type	Loading Type/ Location	Load (%BW)	Task	Task Duration	Measurement Method	Outcome Measurement
Al-Khabbaz, Shimada, and Hasegawa (2008)	19	18–24	M	Normal	Symmetrical	0, 10, 15, 20	Standing	5 s	VICON motion analysis system and surface EMG	Trunk inclination and muscle activity
Brackley, Stevenson, and Selinger (2009)	15	10	M, F	Normal	Symmetrical (high, medium, low)	0, 15	Standing and walking	30 min	Spring-loaded potentiometers	Spinal curvature and trunk inclination
Brzek et al. (2017)	155	7–9	M, F	Normal	Symmetrical	Varied	Standing	—	Pedi-Scoliometer, Dobosiewicz methodology	Spinal curvature and trunk inclination
Chow, Leung, and Holmes (2007)	15	15–16	M	Normal	Symmetrical	0, 10, 15, 20	Standing	—	5 camera motion analysis	Spinal curvature and lumbar repositioning ability
Chow, Ou, Wang, and Lai (2010)	19	10–11	M, F	Normal	Symmetrical anterior and posterior (CG at T7, T12, or L3)	0,15	Standing	—	6 gravitationally referenced accelerometers	Spine curvature and lumbar repositioning ability
Devroey, Jonkers, de Becker, Lenaerts, and Spaepen (2007)	20	20–27	M, F	Normal	Symmetrical (thoracic and lumbar)	0, 5, 10, 15	Standing and walking	1–5 min	6 camera VICON system	Trunk kinematics and muscle activity
Drzal-Grabiec, Snela, et al. (2015)	162	11–13	M, F	Normal	Asymmetrical	0, 10	Standing	—	Photogrammetry	Spinal posture
Drzal-Grabiec, Truszczynska, et al. (2015)	162	11–13	M, F	Normal	Asymmetrical	0, 10	Standing	—	Photogrammetry	Spinal posture
Goh, Thambyah, and Bose (1998)	10	18–21	M	Normal	Symmetrical	0, 15, 30	Walking	—	5 camera motion analysis	L5/S1 joint deformation
Goodgold et al. (2002)	2	9–11	M	Normal	Symmetrical	0, 8.5, 17	Standing, walking, and running	—	Peak 5 motion analysis videography	Trunk inclination
Grimmer et al. (2002)	250	12–18	M, F	Normal	Symmetrical (high, medium, low)	0, 3, 5, 10	Standing	—	Photograph analysis with anatomical markers	Spinal posture
Hong and Brueggemann (2000)	15	10	M	Normal	Symmetrical	0, 10, 15, 20	Walking	~1 min	3-CCD camera and motion analysis	Trunk inclination
Hong and Cheung (2003)	11	9–10	M	Normal	Symmetrical	0, 10, 15, 20	Walking	20 min	Video motion analysis	Trunk inclination

(continued)

TABLE 2: (continued)

Study	Sample Size	Age (Years)	Gender	Backpack Type	Loading Type/ Location	Load (%BW)	Task	Task Duration	Measurement Method	Outcome Measurement
Hong, Fong, and Li (2011)	13	11–13	M	Single strap, athletic bag, and normal backpack	Symmetrical and asymmetrical	0, 10,15, 20	Stairs ascending and descending	—	Video motion analysis	Trunk inclination
Kistner, Fiebert, Roach, and Moore (2013)	62	8–11	M, F	Normal	Symmetrical	0, 10, 15, 20	Standing and walking	6 min	Photograph analysis	Spinal curvature and trunk inclination
Li, Hong, and Robinson (2003)	15	10	M	Normal	Symmetrical	0, 10, 15, 20	Walking	20 min	Video analysis	Trunk inclination
Mackie et al. (2008)	16	13–14	M	Normal	Symmetrical	0, 5, 10, 12.5, 15	Simulated school day	~123 min over 6 days	Video analysis using anatomical markers	Spinal posture
Pascoe, Pascoe, Wang, Shim, and Kim (1997)	10	11–13	M, F	One- and two-strap backpack and one-strap athletic bag	Symmetrical and asymmetrical	—	Standing and walking	—	Video analysis	Spinal curvature and trunk inclination
Ramprasad, Alias, and Raghuveer (2010)	209	12–13	M	Normal	Symmetrical	0, 5, 10, 15, 20, 25	Standing	—	Photograph analysis with anatomical markers	Trunk inclination
Shymon et al. (2014)	15	7–17	M, F	Normal	Symmetrical	0, ~10, ~20	Standing	~10 min	MRI scanner	Spinal curvature and lumbar disk deformation
Singh et al. (2009)	17	7–11	M	Normal	Symmetrical (high, low)	0, 10, 15, 20	Standing and walking	6 min	6 camera motion capture	Trunk inclination
Walicka-Cuprys, Skalska-Izdebska, Rachwal, and Truszczynska (2015)	109	7	M, F	Varied per subject	Varied per subject	Varied per subject	Standing	~50 min	Ultrasonic 3D analysis	Spinal curvature

Note. A typical school backpack is referred to as normal backpack. BW = body weight; M = male; F = female; EMG = electromyography; CCD = charged coupled device; MRI = magnetic resonance imaging; 3D = three-dimensional.

condition after only 1 min (Li, Hong, & Robinson, 2003). Goodgold et al. assessed trunk posture for two male subjects during running under various backpack weights. They found the maximum trunk forward inclination not to change in one subject but decrease ~7° in the other subject when increasing backpack weight from 8.5% to 17.5% of BW. The maximum average for the no-backpack condition was 14.2° (Goodgold et al., 2002).

For ascending stairs (33 steps), the lumbar flexion of 13 male children (average age 12.2 years) was investigated. Increasing the backpack load up to 20% of BW was not found to affect range of lumbar flexion during ascending and descending stairs, but there were >6° larger lumbar range of flexion in ascending versus descending stairs (Hong, Fong, & Li, 2011).

Muscle activity. During standing, an increase in rectus abdominis and obliques activity and a decrease in bilateral muscle activity of the erector spinae longissimus have been reported for a backpack load of 15% of BW when compared with no-backpack condition (Devroey, Jonkers, de Becker, Lenaerts, & Spaepen, 2007).

Using 10 males of mean age 19.9 years, Goh et al. investigated the effects of backpack loading on lower back net moment during walking. They observed that carrying a given backpack load resulted in a nonlinear increase in the L5/S1 joint moment (26.67% for a load of 15% of BW; 64% for a load of 30% of BW) (Goh, Thambyah, & Bose, 1998). Such disproportionate increase in L5/S1 moment suggests a substantial demand on trunk muscles to offset the task demand.

Effects of Backpack Position

Posture and kinematics. In addition to the backpack weight, the position (vertical and horizontal) of the backpack relative to the back affects spine kinematics and kinetics. Contrary to the widespread belief that backpacks should be worn high on the back (Brackley et al., 2009), most studies indicated that children experience the least amount of postural deviations when the backpack is placed low on the back (Brzek et al., 2017; Grimmer, Dansie, Milanese, Pirunsan, & Trott, 2002; Singh & Koh, 2009). Apart from changes in spinal posture under symmetric backpack load, studies reported excessive postural deviation, mainly in the coronal plane, under asymmetric load (i.e., backpack on the left or right shoulder) (Brzek et al., 2017; Singh & Koh, 2009).

For standing posture, a study involving 162 children (82 girls and 80 boys) aged 11 to 13 years found that asymmetric backpack loads compared with no backpack resulted in ~11% reduction in thoracic kyphosis (Drzal-Grabiec, Snela, Rachwal, Podgorska, & Rykala, 2015; Drzal-Grabiec, Truszczynska, et al., 2015). However, none of these studies reported the outcome measures for symmetric loading. Thoracic placement (top of the backpack on the shoulder line) compared with lumbar placement (bottom of the backpack carried just above the spina iliac posterior superior) of backpack was found to be associated with larger pelvic forward rotation (~4°) and larger hip flexion (~3°) (Devroey et al., 2007). Although there was no significant change in lumbar flexion or thoracic rotation for either placement compared with the no-backpack condition during standing, there was a trend that included lumbar extension for thoracic placement and lumbar flexion for lumbar placement. Placement of backpack on thorax versus lumbar spine was found to cause changes in thorax and lumbar posture during walking similar to those observed during upright standing, except for an increase in lumbar flexion and trunk range of motion (Devroey et al., 2007).

Both anterior (front of body) and posterior (back of body) placement of backpack on the trunk resulted in changes in spinal posture, changes that were magnified with increasing backpack load (Chow, Ou, Wang, & Lai, 2010). When the backpack was placed anteriorly with its center of mass located at the T7 spinal level, an increase in pelvic backward tilt (5.5°) was observed. When placed posterior on the trunk, with the backpack's center of mass at the T7, T12, and L3, there were 6.0°, 5.4°, and 3.3° increases in lower lumbar spine flexion, respectively (Chow et al., 2010). Furthermore, for the same order of positions (i.e., T7, T12, and L3), there were significant increases in upper thoracic rotation (4.4°), lower thoracic rotation (2.0°), and upper lumbar flexion (3.0°), respectively (Chow et al., 2010). The smallest change

in spinal posture was observed when the backpack's center of mass was positioned in front and at the T12 level (Chow et al., 2010).

For asymmetric backpack loading when ascending stairs, Hong et al. (2011) reported an increase of ~3° in trunk lateral bending toward the supported side (strap side) and a decrease of ~2.7° in trunk lateral bending of the loaded side compared with stairs ascending with symmetric backpack loading (Hong et al., 2011). A similar pattern of results but with smaller difference between symmetric and asymmetric loading were found during stairs descending (Hong et al., 2011).

Muscle activity. In general, regardless of backpack positioning, there were significant changes, relative to a no-backpack condition, of bilateral trunk muscle activity for walking tasks (Devroey et al., 2007). These included increases in activity of abdominal muscles and decreases in activity of erector spinae without any backpack-induced co-activation (Devroey et al., 2007).

DISCUSSION

The objective of this narrative review was to summarize the findings of studies that have investigated the effects of carrying school backpacks on the lower back of young people. Although narrative reviews serve as useful educational tools, they do not offer a foundation for design of intervention or making clinical decisions (Green, Johnson, & Adams, 2006). When interpreting the results of studies discussed in this review, the readers should keep in mind that the strengths and weaknesses of the reviewed studies were not discussed due to the nature of this narrative review (e.g., as compared with systematic reviews).

Carrying a backpack in upright standing posture was reported to cause a forward inclination of trunk among young individuals that increased with the weight of backpack as well as the height of its center of mass but decreased if backpack was carried in front of the trunk. There was also an exception wherein carrying backpack caused backward inclination of trunk, an exception that likely was due to the type of load used inside the backpack (i.e., sand versus books). A reduction in lumbar lordosis (i.e., flattening) was also reported with increases in the backpack weight and the height of its center of mass. Backpack-induced flattening of lumbar lordosis is consistent with report of decrease in the height of anterior aspect of lumbar disk from an imaging study. It is also consistent with reports of decrease in activity of extensor muscles and increase in activity of abdominals (i.e., reflecting a shift in demand of the task from flexion to extension demand). Carrying backpack in walking compared with upright standing was reported to cause larger backpack-induced changes in trunk kinematics and lumbar posture, changes that further increased with pace and distance of walking. Trunk kinematics and lumbar posture, however, were not found to be affected by backpack when ascending or descending stairs. Finally, carrying backpack on one shoulder was reported to cause asymmetric trunk kinematics mainly due to deviations in the coronal plane.

Abnormal mechanics of the lower back, including excessive forces and deformations, have been shown to directly and indirectly irritate pain-sensitive nerve endings in tissue and cause LBP (Adams, 2004; Marras, 2008; McGill, 2007; White, 1990). The backpack-induced changes in trunk inclination are likely a response to the posterior shift in the upper body center of mass due to addition of backpack mass. Such a posterior shift in the upper body center of mass negatively affects whole body balance that is likely offset by the reported backpack-induced trunk forward inclination (Winter, 2005). The reported changes in the posture of lumbar spine and trunk muscle activity, on the contrary, are likely in response to changes in equilibrium and stability requirements of the spine that are directly related to forces and deformation experienced in the lower back. Specifically, the added load of backpack imposes an extension moment (a flexion moment in the case of anteriorly positioned backpack) on the spine that should be balanced internally by the active and passive responses of lower back tissues to keep the trunk in the desired posture. Alterations in lumbar posture reflect changes in passive contribution of lower back tissues (due to stretching) to spine equilibrium, whereas changes in measured muscle activity indicate alterations in active contribution of lower back tissues (muscles).

Furthermore, increasing the height of backpack center of mass, while may not affect the equilibrium demand of the backpack on the spine, increases stability requirement of carrying backpack that should primarily be balanced by the active muscle responses (Granata & Orishimo, 2001). These changes in lumbar spine equilibrium and stability demands due to carrying backpack lead to substantial increases in spinal loads even under activities that are not physically demanding (e.g., walking). However, the actual impact of backpack on spinal loads (i.e., the resultant of internal tissue responses and external mechanical demand of the task) during daily activities remains unclear. In other words, trunk kinematics, lower back posture, and trunk muscle activities, which are affected by carrying backpack, influence spinal loads but should be coupled with biomechanical models to acquire estimates of backpack-induced changes in loads experienced in spinal tissues. Such future studies could also include investigation of the effects of the magnitude and distribution of backpack weight as well as backpack position and its attachments to trunk on spinal loads.

The risk of fatigue failure of spinal tissues under typical repetitions of daily activities (e.g., 10,000 steps walking) is relatively low for the magnitude of spinal loads experienced during most daily activities. However, the risk of fatigue failure substantially increases with even modest increases in spinal loads associated with carrying a backpack (Brinckmann et al., 1988; Gallagher & Heberger, 2013). To better understand the role of carrying a school backpack on the development of LBP among children, it is therefore important to determine backpack-induced changes in spinal loads due to not only the immediate but also the prolonged effects of carrying a backpack on lower back mechanics. For instance, estimates of spinal loads calculated for one cycle of a repetitive task can be used in combination with fatigue failure models of spinal tissues (Motiwale, Subramani, Kraft, & Zhou, 2018) to acquire insight related to risk of fatigue failure for spinal tissues associated with a given number of repetition (e.g., number of steps) of that task. Such studies will offer an important foundation not only for better design of school backpacks via ergonomics principles (e.g., in terms of load distribution and contact with the trunk) but also for the development of recommendations for durations of carrying backpack that could mitigate the prolonged adverse biomechanical effects of current school backpacks.

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank Beth Axtell for her contribution as technical editor for this review. C.S. and I.S. were supported in part by an award (5R03HD086512-02) from the National Center for Medical Rehabilitation Research (NIH-NICHD) and an award from the Office of the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Orthopaedic Research Program (award #W81XWH-14-2-0144).

KEY POINTS

- Back pain is an increasing concern in young individuals.
- Backpacks have been suggested to play a pathogenic role in developing back pain in adolescents.
- Researchers have observed deviations in trunk kinematics, lumbar posture, and trunk muscle activities while wearing backpacks during activities of daily living.
- Future research on the immediate and prolonged effects of carrying backpack on spinal loads can inform design of intervention aimed at reduced risk of low back injury.

REFERENCES

Adams, M. A. (2004). Biomechanics of back pain. *Acupuncture in Medicine*, *22*, 178–188.

Adams, M. A. (2013). *The biomechanics of back pain* (3rd ed.). Edinburgh, Scotland: Churchill Livingstone.

Al-Khabbaz, Y. S., Shimada, T., & Hasegawa, M. (2008). The effect of backpack heaviness on trunk-lower extremity muscle activities and trunk posture. *Gait & Posture*, *28*, 297–302. doi:10.1016/j.gaitpost.2008.01.002

Brackley, H. M., Stevenson, J. M., & Selinger, J. C. (2009). Effect of backpack load placement on posture and spinal curvature in prepubescent children. *Work*, *32*, 351–360. doi:10.3233/WOR-2009-0833

Brinckmann, P., Biggemann, M., & Hilweg, D. (1988). Fatigue fracture of human lumbar vertebrae. *Clinical Biomechanics, 3*(Suppl. 1), i–S23. doi:10.1016/S0268-0033(88)80001-9

Brzek, A., Dworrak, T., Strauss, M., Sanchis-Gomar, F., Sabbah, I., Dworrak, B., & Leischik, R. (2017). The weight of pupils' schoolbags in early school age and its influence on body posture. *BMC Musculoskeletal Disorders*, *18*(1), Article 117. doi:10.1186/s12891-017-1462-z

Calvo-Munoz, I., Gomez-Conesa, A., & Sanchez-Meca, J. (2013). Prevalence of low back pain in children and

adolescents: A meta-analysis. *BMC Pediatrics, 13*, Article 14. doi:10.1186/1471-2431-13-14

Cholewicki, J., Panjabi, M. M., & Khachatryan, A. (1997). Stabilizing function of trunk flexor-extensor muscles around a neutral spine posture. *Spine*, *22*, 2207–2212.

Chow, D. H., Leung, K. T., & Holmes, A. D. (2007). Changes in spinal curvature and proprioception of schoolboys carrying different weights of backpack. *Ergonomics*, *50*, 2148–2156. doi:10.1080/00140130701459832

Chow, D. H., Ou, Z. Y., Wang, X. G., & Lai, A. (2010). Short-term effects of backpack load placement on spine deformation and repositioning error in schoolchildren. *Ergonomics*, *53*, 56–64. doi:10.1080/00140130903389050

Devroey, C., Jonkers, I., de Becker, A., Lenaerts, G., & Spaepen, A. (2007). Evaluation of the effect of backpack load and position during standing and walking using biomechanical, physiological and subjective measures. *Ergonomics*, *50*, 728–742. doi:10.1080/00140130701194850

Drzal-Grabiec, J., Snela, S., Rachwal, M., Podgorska, J., & Rykala, J. (2015). Effects of carrying a backpack in an asymmetrical manner on the asymmetries of the trunk and parameters defining lateral flexion of the spine. *Human Factors*, *57*, 218–226. doi:10.1177/0018720814546531

Drzal-Grabiec, J., Truszczynska, A., Rykala, J., Rachwal, M., Snela, S., & Podgorska, J. (2015). Effect of asymmetrical backpack load on spinal curvature in school children. *Work*, *51*, 383–388. doi:10.3233/WOR-141981

Gallagher, S., & Heberger, J. R. (2013). Examining the interaction of force and repetition on musculoskeletal disorder risk: A systematic literature review. *Human Factors*, *55*, 108–124. doi:10.1177/0018720812449648

Goh, J. H., Thambyah, A., & Bose, K. (1998). Effects of varying backpack loads on peak forces in the lumbosacral spine during walking. *Clinical Biomechanics*, *13*(1 Suppl. 1), S26–S31.

Goodgold, S., Mohr, K., Samant, A., Parke, T., Burns, T., & Gardner, L. (2002). Effects of backpack load and task demand on trunk forward lean: Pilot findings on two boys. *Work*, *18*, 213–220.

Granata, K. P., & Orishimo, K. F. (2001). Response of trunk muscle coactivation to changes in spinal stability. *Journal of Biomechanics*, *34*, 1117–1123.

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine*, *5*, 101–117. doi:10.1016/S0899-3467(07)60142-6

Grimmer, K., Dansie, B., Milanese, S., Pirunsan, U., & Trott, P. (2002). Adolescent standing postural response to backpack loads: A randomised controlled experimental study. *BMC Musculoskelet Disord*, *3*, 10.

Hong, Y., & Brueggemann, G. P. (2000). Changes in gait patterns in 10-year-old boys with increasing loads when walking on a treadmill. *Gait & Posture*, *11*, 254–259.

Hong, Y., & Cheung, C. K. (2003). Gait and posture responses to backpack load during level walking in children. *Gait & Posture*, *17*, 28–33.

Hong, Y., Fong, D. T., & Li, J. X. (2011). The effect of school bag design and load on spinal posture during stair use by children. *Ergonomics*, *54*, 1207–1213. doi:10.1080/00140139.2011.615415

Kistner, F., Fiebert, I., Roach, K., & Moore, J. (2013). Postural compensations and subjective complaints due to backpack loads and wear time in schoolchildren. *Pediatric Physical Therapy*, *25*, 15–24. doi:10.1097/PEP.0b013e31827ab2f7

Li, J. X., Hong, Y., & Robinson, P. D. (2003). The effect of load carriage on movement kinematics and respiratory parameters in children during walking. *European Journal of Applied Physiology*, *90*, 35–43. doi:10.1007/s00421-003-0848-9

Mackenzie, W. G., Sampath, J. S., Kruse, R. W., & Sheir-Neiss, G. J. (2003). Backpacks in children. *Clinical Orthopaedics and Related Research*, *409*, 78–84. doi:10.1097/01.blo.0000058884.03274.d9

Mackie, H. W., & Legg, S. J. (2008). Postural and subjective responses to realistic schoolbag carriage. *Ergonomics*, *51*, 217–231. doi:10.1080/00140130701565588

Marras, W. S. (2008). *The working back: A systems view*. Hoboken, NJ: John Wiley.

McGill, S. (2007). *Low back disorders* (1st ed.). Champaign IL: Human Kinetics.

Motiwale, S., Subramani, A., Kraft, R. H., & Zhou, X. (2018). A non-linear multiaxial fatigue damage model for the cervical intervertebral disc annulus. *Advances in Mechanical Engineering*, *10*(6). doi:10.1177/1687814018779494

Negrini, S., & Carabalona, R. (2002). Backpacks on! Schoolchildren's perceptions of load, associations with back pain and factors determining the load. *Spine*, *27*, 187–195.

Nicolet, T., Mannion, A. F., Heini, P., Cedraschi, C., & Balague, F. (2014). No kidding: Low back pain and type of container influence adolescents' perception of load heaviness. *European Spine Journal*, *23*, 794–799. doi:10.1007/s00586-014-3213-2

O'Day, K. (2008). Researchers explore backpack connection to back pain. *Biomechanics*, *15*(9), 39–45.

Pascoe, D. D., Pascoe, D. E., Wang, Y. T., Shim, D. M., & Kim, C. K. (1997). Influence of carrying book bags on gait cycle and posture of youths. *Ergonomics*, *40*, 631–641. doi:10.1080/001401397187928

Potvin, J. R., & O'Brien, P. R. (1998). Trunk muscle co-contraction increases during fatiguing, isometric, lateral bend exertions. Possible implications for spine stability. *Spine, 23*, 774–780; discussion 781.

Ramprasad, M., Alias, J., & Raghuveer, A. K. (2010). Effect of backpack weight on postural angles in preadolescent children. *Indian Pediatrics*, *47*, 575–580.

Roth-Isigkeit, A., Thyen, U., Stoven, H., Schwarzenberger, J., & Schmucker, P. (2005). Pain among children and adolescents: Restrictions in daily living and triggering factors. *Pediatrics*, *115*, e152–e162. doi:10.1542/peds.2004-0682

Sheir-Neiss, G. I., Kruse, R. W., Rahman, T., Jacobson, L. P., & Pelli, J. A. (2003). The association of backpack use and back pain in adolescents. *Spine*, *28*, 922–930. doi:10.1097/01.BRS.0000058725.18067.F7

Shymon, S. J., Yaszay, B., Dwek, J. R., Proudfoot, J. A., Donohue, M., & Hargens, A. R. (2014). Altered disc compression in children with idiopathic low back pain: An upright magnetic resonance imaging backpack study. *Spine*, *39*, 243–248. doi:10.1097/BRS.0000000000000114

Singh, T., & Koh, M. (2009). Effects of backpack load position on spatiotemporal parameters and trunk forward lean. *Gait & Posture*, *29*, 49–53. doi:10.1016/j.gaitpost.2008.06.006

Walicka-Cuprys, K., Skalska-Izdebska, R., Rachwal, M., & Truszczynska, A. (2015). Influence of the weight of a school backpack on spinal curvature in the sagittal plane of seven-year-old children. *BioMed Research International*, *2015*, Article 817913. doi:10.1155/2015/817913

Watson, K. D., Papageorgiou, A. C., Jones, G. T., Taylor, S., Symmons, D. P., Silman, A. J., & Macfarlane, G. J. (2002). Low back pain in schoolchildren: Occurrence and characteristics. *Pain*, *97*, 87–92.

White, A. P. M. (1990). *Clinical biomechanics of the spine* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Whittfield, J. K., Legg, S. J., & Hedderley, D. I. (2001). The weight and use of schoolbags in New Zealand secondary schools. *Ergonomics*, *44*, 819–824. doi:10.1080/00140130117881

Winter, D. A. (2005). *Biomechanics and motor control of human movement*. Hoboken, NJ: John Wiley.

Cazmon Suri earned his MSc in biomedical engineering from the University of Kentucky in 2018. His current research interests include biomechanics of human musculoskeletal systems.

Iman Shojaei earned his PhD in biomedical engineering from the University of Kentucky in 2019. His current research interests include computational methods, biomechanics, and structural engineering.

Babak Bazrgari is an associate professor of biomedical engineering at the University of Kentucky. He received his PhD from École Polytechnique de Montréal in 2008.

Date received: January 4, 2018
Date accepted: May 22, 2019

Effects of Fatigue on Balance Recovery From Unexpected Trips

Xingda Qu, **Yongxun Xie**, **Xinyao Hu**, Shenzhen University, China, and **Hongbo Zhang**, Virginia Military Institute, USA

Objective: The objective was to examine how physical fatigue and mental fatigue affected balance recovery from unexpected trips.

Background: Trips are the leading cause for occupational falls that are a multifactorial problem. Recognizing risk factors is the first step in accident control. Fatigue is one of the most common task-related risk factors for occupational falls. Fatigue typically can be divided into physical fatigue and mental fatigue, both of which are common in occupational settings.

Method: One hundred eight young volunteers participated in the experiment. They were evenly divided into three groups: no fatigue group, physical fatigue group, and mental fatigue group. Each participant performed four walking trials on a linear walkway at their self-selected normal speed. The first three trials were normal walking trials. A trip was induced to participants in the fourth walking trial using a metal pole. Balance recovery from unexpected trips was characterized by trunk flexion and first recovery step measures.

Results: Recovery step length was smaller and maximum trunk flexion became larger in the mental fatigue group compared with those in the no fatigue group. Physical fatigue did not significantly affect trunk flexion and first recovery step measures.

Conclusion: Mental fatigue increased the likelihood of loss of balance. Thus, mental fatigue could be a risk factor for trips and falls. To prevent trip-related falls, interventions should be adopted to prevent prolonged exposures to cognitively demanding activities in occupational settings.

Keywords: falls, trips, physical fatigue, mental fatigue, balance recovery

Address correspondence to Xinyao Hu, Institute of Human Factors and Ergonomics, 3688 Nanhai Avenue, Shenzhen University, Shenzhen, Guangdong Province 518060, China; e-mail: huxinyao@szu.edu.cn.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 919–927
DOI: 10.1177/0018720819858794
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Falls occurring while walking over level ground continue to be a major safety concern in occupational settings (Yeoh, Lockhart, & Wu, 2013). Trips are the leading cause for occupational falls, accounting for nearly one fourth of this kind of accident (Amandus, Bell, Tiesman, & Biddle, 2012; Lipscomb, Glazner, Bondy, Guarini, & Lezotte, 2006). To prevent trips and falls, there is a need for improved knowledge on how people react to unexpected trips, especially from a biomechanics perspective. Such knowledge can aid in better understanding of human postural control mechanisms for balance maintenance after unexpected trips.

Occupational falls are a multifactorial problem. Recognizing risk factors is generally considered as the first step in accident control. Fatigue has been reported as a potential risk factor for occupational falls (Hsiao & Simeonov, 2001). Fatigue itself is multidimensional, and typically can be divided into physical fatigue and mental fatigue (Grandjean, 1979), both of which are common in occupational settings.

Physical fatigue is associated with a reduction of strength or an increase in effort required to accomplish the same physical task. Effects of physical fatigue on fall risks have been well examined. Parijat and Lockhart (2008), for example, presented that gait changes due to fatigue induced by repetitive knee joint exertions contributed to increased risk of slip-induced falls. In a model-based simulation study, ankle fatigue was found to compromise postural control and increase fall risks through delayed sensory feedback (Qu, Nussbaum, & Madigan, 2009). It was also reported that after a fatiguing lifting task, male adults had increased postural sway, which is associated with increased fall risks (Bannon, Hakansson, Jakobson, Sundstrup, & Jorgensen, 2018).

Mental fatigue is a subjective feeling of "tiredness" and "lack of energy" due to prolonged exposure to cognitively demanding activities, and associated with a decrease in maximal cognitive performance (Marcora, Staiano, & Manning, 2009). Researchers have found that postural control demands cognitive resources (Woollacott & Shumway-Cook, 2002). Dual task paradigms have been used to examine the relationship between postural control and cognitive demands. In the dual task paradigms, postural control and cognitive tasks are performed simultaneously, and decline of cognitive task performance has been observed (Mohammadi, Mokhtarinia, Jafarpisheh, Kasaeian, & Osqueizadeh, 2018; Redfern, Jennings, Martin, & Furman, 2001). This indicates that postural control interacts with cognitive tasks and shares overlapping cognitive resources. Some researchers also argued that the postural control performance decreased due to the decreased cognitive resources allocated to postural control when cognitive load was applied (Schmid, Conforto, & Lopez, 2007).

Mental fatigue could attenuate cognitive resource allocation to cognitively demanding tasks (Kato, Endo, & Kizuka, 2009). Therefore, mental fatigue could compromise the control of postural balance and result in increased fall risks. However, compared to physical fatigue, mental fatigue has been relatively less studied in terms of its effects on fall risks. Behrens et al. (2018) observed increased gait variability after mental fatigue and further recommended mental fatigue as an intrinsic risk factor for accidental falls. We also examined the effects of mental fatigue on biomechanics of slips and found that mental fatigue led to increased likelihood of slip initiation, poorer slip detection and a more insufficient reactive recovery response to slips (Lew & Qu, 2014a). These findings suggest that mental fatigue is a risk factor for slips and falls.

Although the relationship between fatigue and fall risks has been investigated, there is still lack of empirical evidence regarding how fatigue could affect the risk of falls after unexpected trips. To address this research gap, we aimed to reveal the effects of fatigue on balance recovery from unexpected trips. Both physical fatigue and mental fatigue were examined. We hypothesized that both physical fatigue and mental fatigue would compromise balance recovery from unexpected trips: After being tripped, people with physical fatigue and mental fatigue would have increased step-off time and recovery step duration, decreased recovery step length, and increased trunk flexion.

METHOD

Participants

One hundred and eight young volunteers participated in the experiment. These participants were all healthy and did not have any medical conditions. They were evenly divided into three groups: no fatigue group (baseline), physical fatigue group, and mental fatigue group, using the matched groups design. The variables considered for matching were age, height, and weight. Participants' demographic information is presented in Table 1. This research complied with the tenets of the Declaration of Helsinki and was approved by the institutional review board at Shenzhen University. Informed consent was obtained from each participant. Before signing the informed consent form, participants were informed that an external perturbation might be applied during the walking trial, but they were not informed of the timing and location of its occurrence.

Apparatus

An eight-camera optoelectronic motion capture system (V5; Vicon, Oxford, the United Kingdom) was used to collect the whole-body kinematic data. Reflective markers were placed on the body following the scheme suggested by the Plug-in Gait model. The motion capture system was sampled at the rate of 100 Hz. Walking trials were conducted on a raised linear walkway (12 m long), which is covered by vinyl tile to represent a realistic environmental setting. An overhead harness protection system was used to protect participants from fall impacts onto the ground. The harness was attached to a bearing on a ceiling mounted track by an inelastic rope.

Experimental Procedure

Prior to data collection, participants were instructed to practice walking on the linear platform

TABLE 1: Participants' Demographic Information and Pretrip Gait Parameters

Demographic Information and Pretrip Gait Parapeters	No Fatigue	Physical Fatigue	Mental Fatigue	*p*
Number: *n*	36	36	36	—
Male: *n*	19	19	19	—
Age: *M* (*SD*)	21.7 (1.9) years	21.7 (1.4) years	21.4 (1.8) years	.778
Body weight: *M* (*SD*)	57.9 (12.7) kg	56.6 (9.0) km	57.7 (10.2) km	.855
Height: *M* (*SD*)	167.5 (8.9) cm	169.2 (8.2) cm	168.1 (9.5) cm	.695
Walking speed: *M* (*SD*)	126.5 (20.7) cm/s	126.8 (17.0) cm/s	129.7 (20.0) cm/s	.797
Step length: *M* (*SD*)	67.2 (7.3) cm	67.1 (7.7) cm	68.9 (9.4) cm	.638

Note. P values were obtained from one-way ANOVA for the comparisons among fatigue groups. ANOVA = analysis of variance.

at their self-selected normal speed. After practice, the no fatigue group started walking trials immediately, whereas fatigue groups (i.e., physical fatigue group and mental fatigue group) were instructed to perform fatiguing exercises before starting walking trials. Details about fatiguing protocols were given in the below section.

Each participant performed four walking trials on the linear walkway at their self-selected normal speed. To take attention away from the floor surface during walking, participants were instructed to look straightforward at a whiteboard at the end of the walkway. Meanwhile, the light was dimmed and participants wore dark sunglasses in the experiment. Dimmed lighting and wearing the dark sunglasses allowed participants to see the whiteboard, but prevented participants' awareness of changes near the walking surface. This setting was used to mimick a realistic scenario (i.e., insufficient lighting scenario) where a substantial number of occupational falls have been reported to take place (Hsiao & Simeonov, 2001). The first three trials were normal walking trials. A trip was induced to participants in the fourth walking trial using a metal pole that was placed 10 cm over the ground, and about 7 m away from the starting point along the walking direction (Figure 1). There was a 1-min break between two consecutive walking trials. During the break, participants faced away from the walkway and listened to music, distracting them from possible setup of the trip device. In addition, all the walking trials for a participant could be completed within 5 min to minimize recovery from fatigue.

Fatiguing Protocols

Physical fatigue was induced by a repetitive stand-to-sit exercise. One repetition included standing up from a 45-cm-high armless chair and then sitting down onto the chair again. This procedure was repeated following a metronome at the rate of 40 Hz. Borg's ratings of perceived exertion were reported to be highly related to physical fatigue after short-term high-intensity exercises (Ratel, Duche, & Williams, 2006). Thus, Borg's ratings of perceived exertion were used to assess participants' physical fatigue level. The Borg's 6-20 scale was selected, in which "6" corresponds to *no fatigue at all* and "20" corresponds to *extremely fatigued*. At every 10-s interval during the physical fatiguing exercise, participants were asked to rate their physical fatigue level using the Borg's 6-20 scale. The fatiguing exercise was stopped and fatigue was considered to be induced when participants first gave a rating at or above 17, which corresponds to "very fatigued" on the Borg's 6-20 scale.

The protocol for inducing mental fatigue was similar with our previous study (Lew & Qu, 2014a). In particular, the mental fatigue group was asked to perform an AX-continuous performance test (AX-CPT) for 90 min. In the AX-CPT task, sequences of letters in white were presented on the center of the computer screen with a black background. Participants sitting in front of the computer screen were instructed to press

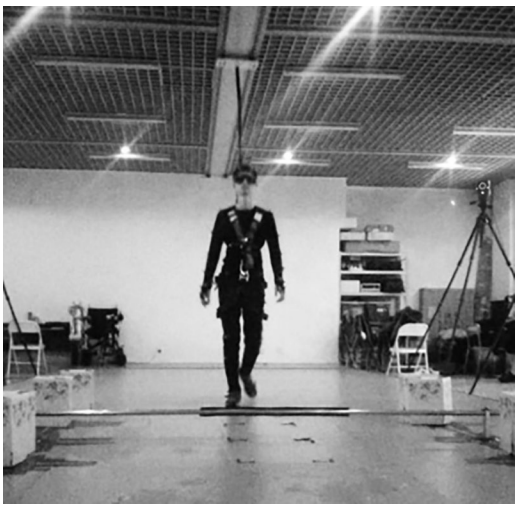


Figure 1. A trip trial.

the right button of the mouse in response to an appearance of the letter "X" following a letter "A" (target trials). Otherwise, they were required to press the left button of the mouse. Each letter appeared on the screen for 300 ms and intervals between two consecutive letters were 1200 ms. Participants were asked to rate their mental fatigue level right before the AX-CPT task (prefatigue), right after the AX-CPT task (postfatigue), and right after the completion of walking trials (posttrip), respectively, using the visual analog scale (VAS; 100 mm in length).

Data Reduction

Three participants in the no fatigue group, six participants in the physical fatigue group, and five participants in the mental fatigue group were not effectively tripped (e.g., one foot stepped onto the trip pole, stopped before the trip pole, or both feet stepped over the trip pole) when the trip pole was present. Data from these participants were excluded from analysis. One participant in the no fatigue group, one participant in the physical fatigue group, and four participants in the mental fatigue group fell right after being tripped without executing a recovery step. Their data were also not used here, as recovery performance could not be assessed. In addition, reflective markers at anterior superior iliac spines (ASIS) from two participants in the no fatigue group and one participant in the mental fatigue group were missing during trip recovery. After reducing these data, 30 trip trials in the no fatigue group, 29 trip trials in the physical fatigue group, and 26 trip trials in the mental fatigue groups were used in the analysis.

Dependent Variables

As trip perturbations typically result in loss of balance in the forward direction, arresting trunk flexion (i.e., controlling the center of mass [COM] backward) and executing appropriate first recovery step (i.e., enlarging the base of support [BOS]) are critical for successful recovery after trips (Honeycutt, Nevisipour, & Grabiner, 2016). Thus, characteristics of trunk flexion and first recovery step were examined. In particular, dependent variables included step-off time, recovery step duration, recovery step length, maximum trunk flexion displacement, and maximum trunk flexion velocity. To determine these dependent variables, moments of trip onset, nontrip foot release, and nontrip foot contact were calculated based on the approach proposed by Pijnappels, Bobbert, and van Dieen (2001) and Lee, Martin, Thrasher, and Layne (2017). Step-off time was defined as the time from trip onset to the release of the nontrip foot. Recovery step duration was the time interval between nontrip foot release and nontrip foot contact. Step length was the sagittal plane distance between the ankle markers of the two legs at the moment of nontrip foot contact. Trunk flexion angle was determined by the orientation of the trunk relative to the vertical. The trunk orientation was defined by the midpoint of the hip joint centers and the midpoint of the shoulder markers (Qu & Yeo, 2011). The hip joint center was calculated as a function of pelvic width, which was defined by the ASIS separation. Specifically, the hip joint center is located 14% of pelvic width medially, 24% of pelvic width posteriorly, and 30% of pelvic width inferiorly relative to the ASIS (Seidel, Marchinda, Dijkers, & Soutas-Little, 1995). Maximum trunk flexion displacement and velocity were calculated for the period corresponding to recovery step duration.

Note that in some trials (seven "no fatigue" trials, four "physical fatigue" trials, and five "mental fatigue" trials), both feet were caught by the pole in sequence. In these trials, the earlier tripped foot was typically the one that executed the recovery step, so it was defined as the

TABLE 2: Dependent Variables in the Two-Feet and One-Foot Trip Trials

Dependent Variables	One-Foot Trips	Two-Feet Trips	*p*
Step-off time (s)	0.21 (0.16)	0.16 (0.12)	.309
Recovery step duration (s)	0.47 (0.12)	0.43 (0.06)	.253
Recovery step length (m)	0.60 (0.21)	0.62 (0.17)	.707
Max trunk flexion (degree)	30.09 (11.51)	28.10 (10.88)	.531
Max trunk flexion velocity (degree/s)	88.76 (45.88)	112.62 (51.65)	.071

"non-trip foot," whereas the latter tripped foot was considered as the "trip foot." No significant differences in dependent variables were found between one-foot trip trials and two-feet trip trials (Table 2). Therefore, the data in these two conditions were pooled when analyzing fatigue effects.

Analysis

The independent variable was "fatigue," which had three levels: no fatigue, physical fatigue, and mental fatigue. As prior-to-trip gait parameters may influence the trip recovery responses (i.e., dependent variables), to increase the statistical power, initial analyses were conducted using a multivariate analysis of covariance (MANCOVA) with walking speed and step length as covariates. In the MANCOVA, the three fatigue conditions were compared using Wilks's lambda test. Walking speed and step length were determined using the data from the normal gait cycle right before the occurrence of trips.

The MANCOVA allowed for determination of fatigue effects on the dependent variables as a whole. If significant fatigue effects were found by the MANCOVA, univariate analysis of covariance (ANCOVA) was then conducted separately for each dependent variable with walking speed and step length as covariates. As we were only interested in comparing postfatigue conditions (i.e., physical fatigue and mental fatigue) to baseline (i.e., no fatigue), post hoc pairwise comparisons were conducted using the Dunnett's test. In addition, VAS scores were compared among the prefatigue, postfatigue, and posttrip conditions using analysis of variance (ANOVA). Similarly, post hoc pairwise comparisons of VAS scores were conducted using the Dunnett's test. Level of significance for all tests was set at $\alpha = .05$.

RESULTS

The MANCOVA results showed significant fatigue effects (Wilks' lambda = 0.761, $p = .019$). In the subsequent ANCOVAs, significant fatigue effects were found in recovery step length and maximum trunk flexion displacement (Table 3). Results from post hoc pairwise comparisons showed that significant differences in recovery step length and maximum trunk flexion displacement existed between the baseline and mental fatigue conditions (Table 3). On average, recovery step length was 22% smaller and maximum trunk flexion was 25% larger in the mental fatigue group compared with those in the no fatigue group (Table 3). Although significant physical fatigue effects were not observed, we noted that p values for the comparisons of the baseline and physical fatigue conditions in recovery step length and maximum trunk flexion displacement were less than 0.10, suggesting physical fatigue effects approach statistical significance (Table 4).

VAS scores were 10.5(11.2), 68.0(16.5), 45.2(19.1), respectively, for the prefatigue, postfatigue, and posttrip conditions. Significant differences were found between prefatigue and postfatigue conditions ($p < .001$), and between prefatigue and posttrip conditions ($p < .001$).

DISCUSSION

The objective of the present study was to examine how physical fatigue and mental fatigue affected balance recovery from unexpected trips. Balance recovery was characterized by trunk flexion and first recovery step measures. Physical

TABLE 3: Summary of the Results From ANCOVAs: Mean (*SD*)

Dependent Variables	No Fatigue	Physical Fatigue	Mental Fatigue	*p* Values for Fatigue	*p* Values for Covariates: Walking Speed	*p* Values for Covariates: Step Lengths
Step-off time (s)	0.21 (0.15)	0.18 (0.13)	0.21 (0.18)	.638	.530	.969
Recovery step duration (s)	0.45 (0.09)	0.48 (0.15)	0.46 (0.09)	.522	.093	.402
Recovery step length (m)	0.68 (0.19)	0.58 (0.18)	0.53 (0.21)	.010*	.765	.610
Max trunk flexion (degree)	24.99 (7.27)	30.81 (12.77)	32.74 (11.84)	.026*	.897	.849
Max trunk flexion velocity (degree/s)	89.33 (38.74)	87.72 (52.34)	102.00 (50.08)	.365	.154	.886

Note. ANCOVA = analysis of covariance.
*Statistical significance ($p < .05$).

TABLE 4: Summary of the Results From the Post Hoc Pairwise Comparisons

Pairwise Comparisons	Recovery Step Length (m): 95% Confidence Interval (I Minus II) Lower Bound	Upper Bound	*p*	Maximum Trunk Flexion (Degree): 95% Confidence Interval (I Minus II) Lower Bound	Upper Bound	*p*
NF (I) vs. PF (II)	–0.013	0.213	.090	–12.163	0.537	.077
NF (I) vs. MF (II)	0.035	0.268	.008*	–14.279	–1.212	.017*

Note. NF = no fatigue; PF = physical fatigue; MF = mental fatigue.
*Statistical significance ($p < .05$).

fatigue that was associated with decreased capacity of force generation was induced by a repetitive stand-to-sit exercise. Stand-to-sit is a closed kinetic chain movement that is able to induce functional fatigue patterns involving multiple lower limb muscle groups. Roos, McGuigan, and Trewartha (2010) presented that insufficient lower limb strength was a contributor to failed balance recovery from trips. Therefore, we hypothesized that physical fatigue would compromise balance recovery performance. Surprisingly, this hypothesis was not well supported by our findings, as no significant physical fatigue effects were observed. In Lew and Qu (2014b), it was found that lower limb fatigue could result in adaptive safer postural control to maintain the likelihood of slip initiation. Similarly, the findings here may suggest that people could adaptively control first recovery step and trunk forward rotation in the lower limb fatigue condition to avoid a fall after unexpected trips.

In addition, we noted that on average, physical fatigue led to 15% shorter recovery step lengths and 23% larger trunk flexion even though these changes were not significant. In fact, differences between the baseline and physical fatigue conditions in recovery step length and maximum trunk flexion displacement approached significance ($p < .10$; Table 4). This indicates that physical fatigue may have effects

on balance recovery performance, but the physical fatigue protocol adopted in the present study may not be sufficiently challenging to cause significant changes in balance recovery parameters. Previous studies have shown that the effects of physical fatigue on balance recovery from external perturbations were dependent on fatiguing exercises (Lew & Qu, 2014b). Meanwhile, physical fatigue levels in the present study were assessed subjectively. Subjective data have the problem of low validity, which may prevent identifying significant physical fatigue effects. Besides, physical fatigue was often measured in an objective way in previous studies (e.g., Qu et al., 2009). It may be difficult to make comparisons between the present study and other studies involving objective physical fatigue measures. Thus, to further understand how physical fatigue affects balance recovery from unexpected trips, different fatiguing protocols in terms of fatiguing exercises and fatigue assessment approaches need to be incorporated into future research.

We noted that VAS scores were significantly smaller in the prefatigue condition compared with those in the postfatigue and posttrip conditions. This suggests that mental fatigue was successfully induced, and participants were in a mental fatigue state when being tripped in the experiment. We found that mental fatigue was associated with smaller recovery step length and larger anterior trunk rotation (i.e., larger trunk flexion). To maintain postural balance, the vertical projection of the whole-body COM should be kept within the limits of stability, which is determined by the BOS (Hof, Gazendam, & Sinke, 2005). Smaller recovery step length indicates smaller BOS, which decreases the limits of stability and increases the likelihood of loss of balance. Increased anterior trunk rotation is also an indicator of poorer balance control as it moves the COM closer to the anterior boundary of BOS. Therefore, it might be concluded that mental fatigue could increase the risk of loss of balance and compromise balance recovery performance after unexpected trips, supporting our initial hypothesis regarding the mental fatigue effects.

Mental fatigue interferes with cognitive resource allocation during cognitively demanding tasks (Kato et al., 2009). As researchers investigating dual-task effects on cognition and postural control have shown that postural control is cognitively demanding (Woollacott & Shumway-Cook, 2002), poorer balance recovery performance in the mental fatigue condition could be attributed to decreased cognitive resources allocated to postural control. In addition, mental fatigue is associated with less activated prefrontal cortex (Holtzer et al., 2017). Prefrontal cortex is the brain area involved in executive functioning that is essential for gait postural control (Leone et al., 2017). Impaired executive functioning could result in poor self-awareness of physical limitations and further lead to inappropriate evaluation of environmental hazards (Yogev-Seligmann, Hausdorff, & Giladi, 2008). This can explain why participants with mental fatigue performed worse at arresting their trunk flexion and executing appropriate first recovery step after unexpected trips.

Mental fatigue was reported to be associated with longer reaction time (Kato et al., 2009). However, we did not find the effects of mental fatigue on step-off time and recovery step duration in the present study. In Kato et al. (2009), reaction time was recorded when participants voluntarily pressed a button in reaction to a visual stimulus. However, trips in our experiment were induced unexpectedly. Thus, participants' reaction to trips was involuntary in nature. Involuntary reactions demand much less cognitive recourses than do voluntary reactions, as they do not have to go to the brain to be cognitively processed. Meanwhile, it was also reported that mental fatigue did not affect maximal muscle activation in the lower limbs (Pageaux, Marcora, & Lepers, 2013). Therefore, we may speculate that participants with mental fatigue were able to react to trips and execute recovery actions as quickly as in the no fatigue condition. As a result, no differences in reaction time and recovery step duration were observed between the mental fatigue and no fatigue groups.

As many trip-related accidents in occupational settings were induced by pole-like objects (Matern & Koneczny, 2007), we used a metal pole to trip participants in our experiment. However, we noted that other trip devices were also used in the investigations of trips and falls. For instance, some researchers used metal bars/

plates to obstruct the toe of the swing foot (Roos, McGuigan, & Trewartha, 2010; Wang, Bhatt, Yang, & Pai, 2012). Mechanical perturbations applied onto the swing foot are different when using different trip devices, which may directly have effects on postural control response after unexpected trips. Therefore, it is of interest to examine fatigue-related balance recovery from unexpected trips with the application of other trip devices in future research.

The major contribution of this study is that it revealed the effects of fatigue on balance recovery from unexpected trips. Both physical fatigue and mental fatigue were examined. Significant physical fatigue effects were not observed, but we found that mental fatigue increased the likelihood of loss of balance. These findings suggest that mental fatigue could be a risk factor for trips and falls. To prevent trip-related falls, interventions should be adopted to prevent prolonged exposures to cognitively demanding activities in occupational settings.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (Grant 31570944 & 11702175), Natural Science Foundation of Guangdong Province of China (Grant 2015A030313553 & 2016A030310068), and Shenzhen Science, Technology, and Innovation Council (Grant JCYJ20150525092940994).

KEY POINTS

- Physical fatigue was not found to have effects on balance recovery from unexpected trips.
- Mental fatigue increased the likelihood of loss of balance.
- Mental fatigue could be a risk factor for trips and falls.
- Prolonged exposures to cognitively demanding activities should be avoided in occupational settings.

ORCID iD

XingdaQu https://orcid.org/0000-0003-1764-0357

REFERENCES

Amandus, H., Bell, J., Tiesman, H., & Biddle, E. (2012). The epidemiology of slips, trips, and falls in a helicopter manufacturing plant. *Human Factors*, *54*, 387–397. doi:10.1177/0018720811403140

Bannon, H. M., Hakansson, N. A., Jakobson, M. D., Sundstrup, E., & Jorgensen, M. J. (2018). The effects of a fatiguing lifting task on postural sway among males and females. *Human Movement Science*, *59*, 193–200. doi:10.1016/j.humov.2018.03.008

Behrens, M., Mau-Moeller, A., Lischke, A., Katlun, F., Gube, M., Zschorlich, V., . . . Weippert, M. (2018). Mental fatigue increases gait variability during dual-task walking in old adults. *Journal of Gerontology, Series A: Biological Sciences & Medical Sciences*, *73*, 792–797. doi:10.1093/gerona/glx210

Grandjean, E. (1979). Fatigue in industry. *British Journal of Industrial Medicine*, *36*, 175–186. Retrieved from https://www.jstor.org/stable/27723358

Hof, A. L., Gazendam, M. G. J., & Sinke, W. E. (2005). The condition for dynamic stability. *Journal of Biomechanics*, *38*, 1–8. doi:10.1016/j.jbiomech.2004.03.025

Holtzer, R., Yuan, J., Verghese, J., Mahoney, J. R., Izzetoglu, M., & Wang, C. (2017). Interactions of subjective and objective measures of fatigue defined in the context of brain control of locomotion. *The Journals of Gerontology: Series A*, *72*, 417–423. doi:10.1093/gerona/glw167

Honeycutt, C. F., Nevisipour, M., & Grabiner, M. D. (2016). Characteristics and adaptive strategies linked with falls in stroke survivors from analysis of laboratory-induced falls. *Journal of Biomechanics*, *49*, 3313–3319. doi:10.1016/j.jbiomech.2016.08.019

Hsiao, H., & Simeonov, P. (2001). Preventing falls from roofs: A critical review. *Ergonomics*, *44*, 537–561. doi:10.1080/00140130110034480

Kato, Y., Endo, H., & Kizuka, T. (2009). Mental fatigue and impaired response processes: Event-related brain potentials in a Go/NoGo task. *International Journal of Psychophysiology*, *72*, 204–211. doi:10.1016/j.ijpsycho.2008.12.008

Lee, B. C., Martin, B. J., Thrasher, T. A., & Layne, C. S. (2017). The effect of vibrotactile cuing on recovery strategies from a treadmill-induced trip. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*, 235–243. doi:10.1109/TNSRE.2016.2556690

Leone, C., Feys, P., Moumdjian, L., D'Amico, E., Zappia, M., & Patti, F. (2017). Cognitive-motor dual-task interference: A systematic review of neural correlates. *Neuroscience & Biobehavioral Reviews*, *75*, 348–360. doi:10.1016/j.neubiorev.2017.01.010

Lew, F. L., & Qu, X. (2014a). Effects of mental fatigue on biomechanics of slips. *Ergonomics*, *57*, 1927–1932. doi:10.1080/00140139.2014.937771

Lew, F. L., & Qu, X. (2014b). Effects of multi-joint muscular fatigue on biomechanics of slips. *Journal of Biomechanics*, *47*, 59–64. doi:10.1016/j.jbiomech.2013.10.010

Lipscomb, H. J., Glazner, J. E., Bondy, J., Guarini, K., & Lezotte, D. (2006). Injuries from slips and trips in construction. *Applied Ergonomics*, *37*, 267–274. doi:10.1016/j.apergo.2005.07.008

Marcora, S. M., Staiano, W., & Manning, V. (2009). Mental fatigue impairs physical performance in humans. *Journal of Applied Physiology*, *106*, 857–864. doi:10.1152/japplphysiol.91324.2008

Matern, U., & Koneczny, S. (2007). Safety, hazards and ergonomics in the operating room. *Surgical Endoscopy*, *21*, 1965–1969. doi:10.1007/s00464-007-9396–4

Mohammadi, S., Mokhtarinia, H. R., Jafarpisheh, A. S., Kasaeian, A., & Osqueizadeh, R. (2018). Investigating the effects of different working postures on cognitive performance. *Archives of Rehabilitation*, *18*, 268–277. doi:10.21859/JREHAB.18.4.1

Pageaux, B., Marcora, S. M., & Lepers, R. (2013). Prolonged mental exertion does not alter neuromuscular function of the knee extensors. *Medicine & Science in Sports & Exercise*, *45*, 2254–2264. doi:10.1249/MSS.0b013e31829b504a

Parijat, P., & Lockhart, T. E. (2008). Effects of lower extremity muscle fatigue on the outcomes of slip-induced falls. *Ergonomics*, *51*, 1873–1884. doi:10.1080/00140130802567087

Pijnappels, M., Bobbert, M. F., & van Dieen, J. H. (2001). Changes in walking pattern caused by the possibility of a tripping reaction. *Gait & Posture*, *14*, 11–18. doi:10.1016/S0966-6362(01)00110–2

Qu, X., Nussbaum, M. A., & Madigan, M. L. (2009). Model-based assessments of the effects of age and ankle fatigue on the control of upright posture in humans. *Gait & Posture*, *30*, 518–522. doi:10.1016/j.gaitpost.2009.07.127

Qu, X., & Yeo, J. C. (2011). Effects of load carriage and fatigue on gait characteristics. *Journal of Biomechanics*, *44*, 1259–1263. doi:10.1016/j.jbiomech.2011.02.016

Ratel, S., Duche, P., & Williams, C. A. (2006). Muscle fatigue during high-intensity exercise in children. *Sports Medicine*, *36*, 1031–1065. doi:10.2165/00007256-200636120-00004

Redfern, M. S., Jennings, J. R., Martin, C., & Furman, J. M. (2001). Attention influences sensory integration for postural control in older adults. *Gait & Posture*, *14*, 211–216. doi:10.1016/S0966-6362(01)00144–8

Roos, P. E., McGuigan, M. P., & Trewartha, G. (2010). The role of strategy selection, limb force capacity and limb positioning in successful trip recovery. *Clinical Biomechanics*, *25*, 873–878. doi:10.1016/j.clinbiomech.2010.06.016

Schmid, M., Conforto, S., & Lopez, L. (2007). Cognitive load affects postural control in children. *Experimental Brain Research*, *179*, 375–385. doi:10.1007/s00221-006-0795-x

Seidel, G. K., Marchinda, D. M., Dijkers, M., & Soutas-Little, R. W. (1995). Hip joint center location from palpable bony landmarks—A cadaver study. *Journal of Biomechanics*, *28*, 995–998. doi:10.1016/0021-9290(94)00149-X

Wang, T. Y., Bhatt, T., Yang, F., & Pai, Y. C. (2012). Adaptive control reduces trip-induced forward gait instability among young adults. *Journal of Biomechanics*, *45*, 1169–1175. doi:10.1016/j.jbiomech.2012.02.001

Woollacott, M., & Shumway-Cook, A. (2002). Attention and the control of posture and gait: A review of an emerging area of research. *Gait & Posture*, *16*, 1–14. doi:10.1016/S0966-6362(01)00156–4

Yeoh, H. T., Lockhart, T. E., & Wu, X. F. (2013). Non-fatal occupational falls on the same level. *Ergonomics*, *56*, 153–165. doi:10.1080/00140139.2012.746739

Yogev-Seligmann, G., Hausdorff, J. M., & Giladi, N. (2008). The role of executive function and attention in gait. *Movement Disorders*, *23*, 329–342. doi:10.1002/mds.21720

Xingda Qu is a professor with the Institute of Human Factors and Ergonomics at Shenzhen University. He received his PhD degree in human factors engineering and ergonomics from Virginia Tech in 2008, and his research interests lie in gait and balance, injury prevention, and human–computer interaction.

Yongxun Xie was a master student with the Institute of Human Factors and Ergonomics at Shenzhen University. His thesis research focused on investigating postural control mechanisms during recovery from unexpected trips.

Xinyao Hu is an assistant professor with the Institute of Human Factors and Ergonomics at Shenzhen University. He received his PhD degree in human factors engineering and ergonomics from Nanyang Technological University in 2014, and his research interests lie in gait analysis, fall prevention, and sports biomechanics.

Hongbo Zhang is an assistant professor with the Department of Computer and Information Sciences at Virginia Military Institute, the United States. His current interests include human and computer interaction, human motion control, human motion 3D reconstruction, and human collaboration.

Date received: September 13, 2018
Date accepted: May 26, 2019

A Curvilinear Effect of Mental Workload on Mental Effort and Behavioral Adaptability: An Approach With the Pre-Ejection Period

Charlotte Mallat, **Julien Cegarra**, **Christophe Calmettes**, and **Rémi L. Capa**,
INU Champollion, Albi, France

Objective: We tested Hancock and Szalma's mental workload model, which has never been experimentally validated at a global level with the measure of the pre-ejection period (PEP), an index of beta-adrenergic sympathetic impact.

Background: Operators adapt to mental workload. When mental workload level increases, behavioral and physiological adaptability intensifies to reduce the decline in performance. However, if the mental workload exceeds an intermediate level, behavioral and physiological adaptability will decrease to protect individuals from excessive perturbations. This decrease is associated with a change in behavioral strategies and disengagement.

Method: The experimental task was a modified Fitts' task used in Hancock and Caird. Five levels of task difficulty were computed. Behavioral and physiological adaptability was indexed by the performance with speed–accuracy trade-off and PEP reactivity.

Results: A curvilinear effect of task difficulty on PEP reactivity was significant, with high reactivity at the intermediate level but low reactivity at other levels. We observed a linear effect of task difficulty on error rate and a curvilinear effect on movement time. A decline in performance was noted up to the intermediate level, with a speed–accuracy trade-off above this level showing a faster movement time.

Conclusion: We observed for the first time behavioral and physiological adaptability as a function of mental workload.

Application: The results have important implications for the modeling of mental workload, particularly in the context of the performance-sensitive domain (car driving and air traffic control). They can help guide the design of human–computer interaction to maximize adaptive behavior, that is, the "comfort zone."

Keywords: workload, effort, performance, cardiovascular reactivity

Address correspondence to Charlotte Mallat, Laboratoire Sciences de la Cognition, Technologie, Ergonomie (SCoTE—EA 7420), Université de Toulouse, INU Champollion, Place de Verdun, 81012 Albi Cedex 9, France; e-mail: charlotte.mallat@univ-jfc.fr. Rémi L. Capa, Laboratoire Sciences de la Cognition, Technologie, Ergonomie (SCoTE—EA 7420), Université de Toulouse, INU Champollion, Place de Verdun, 81012 Albi Cedex 9, France; e-mail: remi.capa@univ-jfc.fr.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 928–939
DOI: 10.1177/0018720819855919
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Flying an aircraft varies in difficulty depending on situations. The *task demand* can differ substantially when considering time pressure, task complexity, or the performance levels imposed on the operator (Gopher & Donchin, 1986).

Indeed, an operator has to consider the task demand along with his or her characteristics, such as previous experience with the task (i.e., training). Therefore, when evaluating *mental workload* in performing a task, one has to consider the interaction between characteristics of the task, the operator, and the environment (Hart & Staveland, 1988; Wickens, 2008).

While coping with task demand, operators mobilize cognitive resources, which are drawn from a finite pool of resources. In an "energetic" paradigm, one should consider that operators mobilize their resources dynamically and voluntarily (Hockey, 1997; Kahneman, 1973). In our study, we focused on the relationship between mental workload, mental effort, and performance. We used a state-of-the-art cardiovascular measurement of mental effort, the pre-ejection period (PEP), to test our effort-related hypothesis.

The Relationship Between Mental Workload, Mental Effort, and Performance

The effect of mental workload on mental effort and performance is of practical significance, and numerous studies in ergonomics have focused on these relationships. Concerning the effect of mental workload on performance, the insufficient stimulation is known to lead to underload, boredom, and decreased performance (Brookhuis, De Waard, & Fairclough, 2003; Hancock, 2013). Conversely, overload is also known to decrease task performance (Cegarra & Chevalier, 2008).

At the same time, operators are not passive to mental workload increase and as such, they must strategically manage their workload (Moray, Dessouky, Kijowski, & Adapathya, 1991). Strategical management is reflected by changes in strategies and/or performance criteria (Parasuraman & Hancock, 2001). Operators might also differ in their task engagement. Different authors have suggested that an increase in task demand (mental workload) leads to a superior mobilization of mental effort (Young, Brookhuis, Wickens, & Hancock, 2015). When the task demand exceeds the upper limit of resources, one might observe a strong decrease in effort and performance (Cegarra & Hoc, 2006; Wickens, 2001). These relationships between mental workload, performance, and mental effort are well illustrated in the mental workload model of Hancock and Chignell (1988), Hancock and Szalma (2006, 2008), Hancock and Caird (1993), and Hancock and Warm (1989).

A Curvilinear Relationship of Mental Workload With Behavioral and Physiological Adaptability

Hancock and colleagues considered a curvilinear relationship of mental workload with behavioral and physiological adaptability. In this model (Figure 1), behavioral and physiological adaptability varies as a function of stress level. Hancock and Szalma (2006) explicitly recognized that for operators, the primary source of stress is task demand, an important aspect of mental workload. In many stressful situations, operators adapt to their environments. In agreement with Hancock and Szalma (2006), the effort is the cost of resolution and represents the force that acts to oppose the mental workload driven by increasing task demand. The effort is defined here as an adaptive strategy of the operator. Adaptive or compensatory processes depend on the response strategies of a number of body structures that regulate the effects of task demand. Adaptation occurs at a behavioral and physiological level. Psychological adaptability is closely tied to an operator's attentional resource capacity and refers to the behavioral responses, whereas physiological adaptability is related to traditional representations of homeostatic adjustment (Hancock & Warm, 1989). A curvilinear effect of task demand on behavioral and physiological adaptability (Figure 1) illustrates this adaptation.

The level of task demand can vary on a continuum from extremely low to extremely high (i.e., hypostress for very low task demand and hyperstress for very high task demand), and it refers to situations of underload and overload. At a low level of task demand, behavioral and physiological adaptability is low. However, behavioral and physiological adaptability is predicted to increase proportionally with the level of task demand up to the intermediate level of task demand. Thus, from very low task demand level up to intermediate task demand level, an increase in behavioral and physiological adaptability represents a compensatory mechanism to protect performance or slow down the decline in performance. However, at the intermediate level of task demand, behavioral and physiological adaptability is stable, and operators are in a comfort zone of maximum adaptability. If task demand progresses toward overload (very high task demand), then the level of behavioral and physiological adaptability decreases. According to the Hancock and Szalma's (2006) model, when task demand increases and approaches overload, behavioral adaptive strategies can protect the operator from large or excessive perturbations and obviate the need to engage in a costly activity.

This model has been partially validated, although never with both behavioral and physiological measures (Hancock & Szalma, 2006), possibly because a physiological index of mental effort has not been well defined yet.

Physiological Reactivity Related to Mental Effort

Physiological measures are frequently used to classify the level of mental effort (Borghetti, Giametta, & Rusnock, 2017; Marinescu et al., 2018). Different physiological measures, such as skin conductance, temperature, heart rate (HR), or pupil dilatation, are frequently used to establish the level of sympathetic nervous system activity (Capa, Audiffren, & Ragot, 2008). Wright (1996) integrated Motivational Intensity Theory (MIT; Brehm & Self, 1989) with the active coping approach (Obrist, 1981), showing

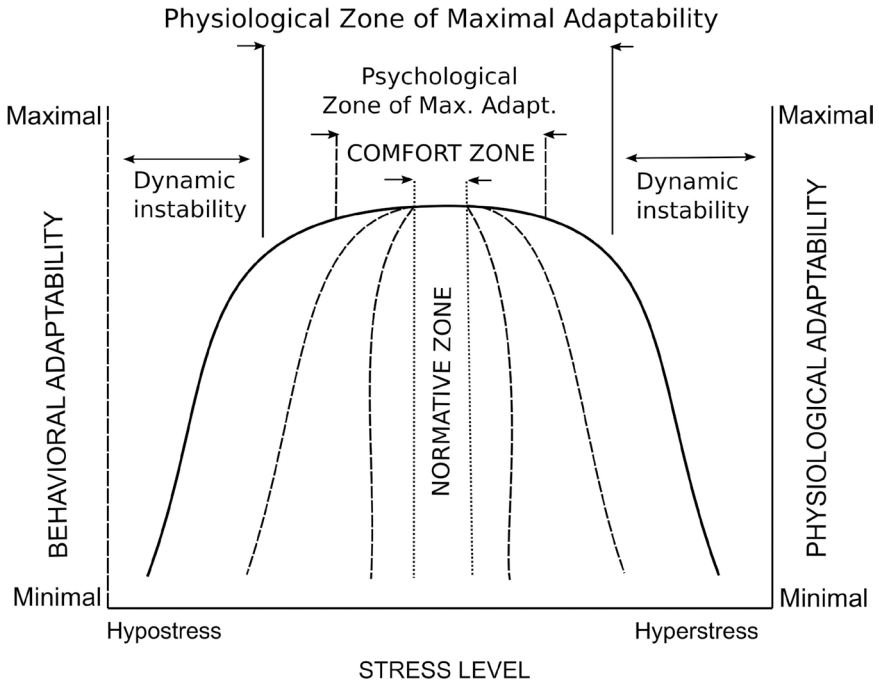


Figure 1. Curvilinear model of the effect of stress level on behavioral and physiological adaptability. The primary source of stress is task demand.
Source. Adapted from Hancock and Szalma (2006, p. 158).

that beta-adrenergic sympathetic nervous system activity on the heart is an indicator of effort. Following this approach, recent studies have highlighted the time interval between ventricular excitation and the opening of the heart's left ventricular valve, or PEP (Berntson, Lozano, Chen, & Cacioppo, 2004), as a more reliable and valid index of mental effort mobilization. PEP is an index of beta-adrenergic sympathetic impact (Benschop et al., 1994). It responds to variations in task difficulty (Richter, Friedrich, & Gendolla, 2008) in a large variety of sensory-cognitive tasks (Richter, Gendolla, & Wright, 2016). Furthermore, PEP is a direct indicator of changes in the myocardial contractility force (Sherwood, Dolan, & Light, 1990). Increased myocardial sympathetic activity increases contraction force and decreases PEP. Due to the systematic effect of cardiac contractility on cardiac output (the volume of blood pumped by the ventricles per minute), several studies have also used systolic blood pressure (SBP) to monitor effort (see Richter et al., 2016; Wright & Kirby, 2001, for reviews). However, SBP is influenced by peripheral vascular resistance and is less sensitive to mental effort investment (Levick, 2003). Still, other studies have relied on HR as an indicator of effort (e.g., Eubanks, Wright, & Williams, 2002), which is influenced by both sympathetic and parasympathetic activity. Increased HR may be due to increased sympathetic activity, but it can also be the result of the decreased parasympathetic activity. Moreover, at low-intensity levels, increases in cardiac activity are due mainly to parasympathetic withdrawal (Victor, Seals, Mark, & Kempf, 1987). Increased HR is influenced by increased sympathetic activity only at high-intensity levels. Among these indices, PEP is the most reliable measure of effort mobilization because it can mirror beta-adrenergic sympathetic activity directly (Kelsey, 2012). Nevertheless, PEP should always be assessed together with HR and blood pressure to monitor possible preload (ventricular filling) or afterload (arterial pressure) effects on PEP (Sherwood et al.,

1990). If decreases in PEP are not accompanied by simultaneous decreases in HR or blood pressure, we can attribute them to beta-adrenergic sympathetic activity.

Because the physiological adaptability described in the mental workload model of Hancock and Szalma (2006) refers to general processes of resource mobilization and not to specific processes of resource mobilization generally measured with electroencephalography or functional magnetic resonance imaging, the PEP is a good candidate to empirically test this model.

Overview of the Experiment and Hypotheses

Our main hypothesis was examined in a modified Fitts' task with shrinking targets. Five levels of task difficulty were found to increase linearly (very easy, easy, intermediate, very difficult, impossible). In agreement with Hancock and Szalma's model, our predictions were as follows. When task demand increases from very low up to intermediate level, then a stronger behavioral and physiological adaptability should occur to reduce the decline in performance. We expected to observe this decline (i.e., an increase of error rate and movement time [MT]) as well as a stronger PEP reactivity related to the mental effort. In agreement with Hockey (1997) and Kahneman (1973), when task demand increases, effort mobilization reduces the decline in performance and helps individuals stay-on-task.

However, if task demand exceeds the intermediate level, then a decrease in behavioral and physiological adaptability should occur to protect individuals from large or excessive perturbations and prevent them from engaging in a costly activity. This decrease in behavioral physiological adaptability would induce a higher error rate and lower PEP reactivity related to the mental effort (Figure 1). Behavioral adaptability should be on the MT (Hoffmann, 2011). To optimize the emergence of behavioral adaptability, the target in the present study will vanish when it reaches one pixel. In fact, time pressure placed on the participants would be high and participants would modify the behavior by decreasing the MT. If the target size continues to decrease (Hancock & Caird, 1993; Johnson & Hart, 1987) but never vanishes, then the movement could be made in a more relaxed manner, and the MT would increase with task difficulty (Hoffmann, Chan, & Dizmen, 2013). Hoffmann (2011) and Hoffmann et al. (2013) observed that when task difficulty increased, the participants moved faster to capture the target prior to it disappearing.

To sum up, we postulated a curvilinear effect of task demand on PEP reactivity, with high PEP reactivity at the intermediate level of task demand and low PEP reactivity at other levels. Concerning the behavioral aspect, we proposed a linear effect of task demand on the error rate. We hypothesized that MT would increase up to the intermediate level and then decrease. In fact, we expected to observe a curvilinear effect of task demand on MT and a speed–accuracy trade-off when task demand exceeds the intermediate level.

METHOD

Participants and Design

One hundred ten postgraduate students at the National University Institute Champollion volunteered to participate. All participants were right-handed and between 18 to 30 years old. All participants had a normal or corrected-to-normal vision and no history of mental, cognitive, cardiac or neurological disorder, substance abuse, or taking psychoactive medications. They were randomly assigned to one of the five difficulty conditions. The distribution of women and men was balanced between experimental conditions, with 17 women and five men per condition.

Procedure

First, the participants read and signed the informed consent agreement and subsequently answered a few biographical questions (Matthews, Altman, Campbell, & Royston, 1990). Second, the electrocardiogram (ECG) and thoracic impedance cardiogram (ICG) electrodes were installed. The signal was visually verified and validated on the AcqKnowledge software. Third, the participants completed the rest period. A video of a chimney fire was presented on the computer screen for 8 min. After that period, the participants received the task instructions. Fourth, they performed 75 trials of training to the experimental task. Fifth, they performed the

experimental task at one of the five difficulty levels (according to the random assignment). Finally, they completed the perceived difficulty scale.

A Modified Fitts' Task With Shrinking Targets

Five levels of difficulty. The experimental task was based on the pointing task of Johnson and Hart (1987), which is a modified Fitts' task (Hancock & Caird, 1993; Hoffmann, 2011). First, we computed different levels of task difficulty in accordance with the Fitts' law (Fitts, 1954).The Fitts' Index of Difficulty (ID) is a function of the amplitude (A) and the size of the target (W). It is expressed in pixels with $ID = \log_2(2A / W)$. The formula indicates that a fast and accurate movement increases in difficulty with decreasing size of a target and increasing amplitudes. In the present study, we selected five IDs (2, 3, 4, 5, and 6) corresponding to five difficulty levels (very easy, easy, intermediate, very difficult, impossible) with a linear increment.

Consistent with Hancock and Caird (1993), we manipulated the perceived effective time for action with a target shrink time (*Ts*). More precisely, the shrink time reflected the time after which the target is half of its current size. Such time was determined by the Fitts' law calculation developed by Hoffmann (2011). More precisely, the shrink time corresponded to the initial target's size (W_0) divided by the speed of targets reduction (V_{sc}) : $T_S = w_0 / V_{sc}$. The target shrinks until it reaches one pixel.

To ascertain that the five selected IDs produce a linear performance decrement with the shrinking targets, we conducted a pretest with 10 participants. In this pretest, we manipulated, in agreement with Hoffmann (2011), different *Ts* (100, 200, 400, 800, and 1600 ms) based on our five difficulty levels, and we analyzed error rate as an index of difficulty. Our analyses revealed that the error rate increased linearly according to the levels of difficulty, that is, IDs that represent task demand. The most notable increase was observed at 200 ms. This value was selected for the experiment.

Experimental task. Each trial had two amplitudes of movement, 50 and 67.7 mm, which were counterbalanced across levels. The amplitudes were calculated with Fitts' law and in accordance with our IDs. Two amplitudes were used to minimize the learning effect of the participants. Moreover, as the two amplitudes of the movement were counterbalanced across task demand conditions, any difference in PEP reactivity would indicate a change in mental effort mobilization rather than a change in physical effort.

The initial target sizes for selected IDs were 96, 48, 24, 12, and 6 pixels with an amplitude of 50 mm and 128, 64, 32, 16, 8 pixels with an amplitude of 67.7 mm, respectively.

The experiment consisted of 150 trials, and each trial was divided into four successive screens (Figure 2). During the preparation phase, only the start position was presented. In the execution phase, the pointer and the targets appeared. The participants were previously instructed to move and click on the targets with the mouse as quickly and precisely as possible. Displacements were limited to the horizontal plane. When a target was clicked, it disappeared while the start position remained on the screen. Finally, the "return" phase was an empty screen. Each trial lasted 5,700 ms.

Cardiovascular Measures and Analyses

Cardiovascular measures (PEP, HR, SBP, diastolic blood pressure [DBP]) were collected during the rest period (8 min) and during the experimental task (15 min). PEP (in milliseconds [ms]) and HR (in beats per minute [bpm]) were measured continuously using a Biopac MP160 system that sampled ECG and ICG signals at 2000 Hz. For the ECG, three electrodes were placed on the right and left shoulder and down to the end of the coasts. For the ICG, electrodes were placed on the right and left sides of the base of participants' neck and on the left and right middle axillary line at the height of the xiphoid. A digital automatic blood pressure monitor (OMRON-MIT Elite) was used to measure arterial pressure (SBP and DBP) in millimeters of mercury (mmHg). The blood pressure cuff was placed over the brachial artery above the elbow of the participants' nondominant arm and was automatically inflated at 1-min intervals. Neither participants nor experimenter were aware of the values during the experiment.

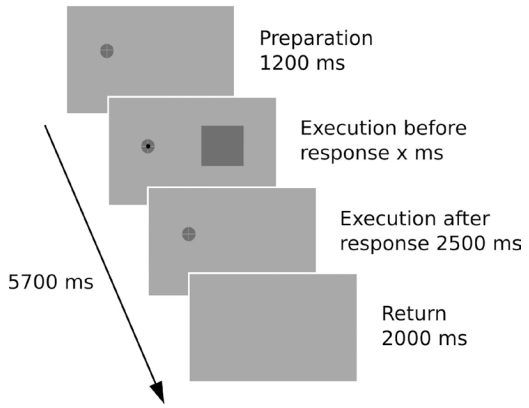


Figure 2. Successive screens displayed for one trial of the pointing task with shrinking targets.

PEP and HR signals were analyzed with the AcqKnowledge 5.0 software. PEP was determined as the time interval (in ms) between R-onset and B-point (Berntson et al., 2004). The R-onset and the B-point were automatically scored for each artifact-free ensemble average. The B-point location was estimated based on locations of R and C points according to a polynomial equation proposed by Lozano et al. (2007), and it was visually inspected and corrected, if necessary (Sherwood et al., 1990).

Rest period. The rest period lasted 8 min. During this period, PEP and HR measures were continuously recorded and averaged for each minute, resulting in eight means. Eight arterial pressure measures (SBP and DBP) were taken at 1 min intervals, resulting in eight measures. Baseline scores were averaged over the three successive lowest means (i.e., PEP and HR) or measures (i.e., SBP and DBP). Cronbach's α values were .99 for PEP baseline, .99 for HR baseline, .98 for SBP baseline, and .93 for DBP baseline scores.

Experimental task. The experimental task lasted 15 min. PEP and HR scores were continuously recorded and averaged for each minute, resulting in 15 means. Arterial pressure (SBP and DBP) was measured after every 50-trial block during the experimental task (50th, 100th, and 150th trial), resulting in three measures. Cardiovascular measures during the experiment had high internal consistencies. Cronbach's α values were .99 for PEP, .99 for HR, .94 for SBP, and .66 for DBP. The arithmetic mean of the three arterial pressure measures during task performance served as an experimental task score. PEP and HR experimental task scores were based on the measures obtained during the 15 min of the experimental task.

Cardiovascular reactivity. Each cardiovascular reactivity score was calculated by subtracting the baseline scores of the rest period from the experimental task scores of the experimental task (Llabre, Spitzer, Saab, Ironson, & Schneiderman, 1991).

Behavioral Measures

The performance was measured by error rate (in %) and MT (in ms). MT is the time it took the participants to click on the target once the movement has been initiated. Only the MTs of the correct responses were examined. Because manipulating the Fitts' law had a large and systematic effect on MT and a relatively small effect on reaction time (Fitts & Peterson, 1964), we focused on MT.

Subjective Measures

The French version of the Eccles and Wigfield's (1995) scale was used to evaluate perceived difficulty. This scale has good psychometric properties and is sensitive to differences in difficulty levels (Capa et al., 2008). It comprises four items measured on a scale from 1 (*very easy*) to 5 (*impossible*) (What is the difficulty level of this task for you?, Compared with difficulties experienced by other participants in this study, this task is. . ., Compared with the difficulty of most other tasks in daily life, this task is. . ., How hard is this task for you?).

Statistical Analysis

Based on our theoretical assumptions, the curvilinear effect of task demand on PEP reactivity was analyzed with a curvilinear polynomial contrast. Quadratic contrast weights were 2, –1, –2, –1, 2 for the difficulty levels from very easy to impossible. Concerning behavior analyses, we tested the linear effect of difficulty levels on error rate with a linear polynomial contrast. Linear contrast weights were –2, –1, 0, 1, 2 for the difficulty levels from very easy to impossible. Subsequently, we tested the curvilinear effect of

TABLE 1: Means (Standard Errors) of Cardiovascular Baseline Scores

	Difficulty Levels				
	Very Easy	Easy	Intermediate	Very Difficult	Impossible
PEP	110.80 (2.77)	107.76 (1.57)	110.05 (2.26)	111.29 (2.11)	108.45 (1.93)
HR	79.95 (2.38)	80.82 (2.52)	82.09 (2.62)	75.96 (2.61)	81.68 (2.26)
SBP	110.86 (1.93)	113.38 (1.57)	112.98 (2.04)	111.62 (2.41)	115.05 (1.68)
DBP	74.09 (1.79)	73.80 (1.77)	72.48 (1.67)	74.36 (2.02)	73.26 (1.82)

Note. $N = 22$ for all cells. The pre-ejection period was measured in milliseconds (ms) and heart rate in beats per minute (bpm). Systolic blood pressure and diastolic blood pressure were measured in millimeters of mercury (mmHg). PEP = pre-ejection period; HR = heart rate; SBP = systolic blood pressure; DBP = diastolic blood pressure.

task difficulty on MT with curvilinear polynomial contrasts. Quadratic contrast weights were 1 for very easy, –1 for easy, –1 for intermediate, and 1 for very difficult. As the error rate in the impossible condition would be close to 100%, very few data would be available for the impossible level of difficulty of MT. Finally, we analyzed subjective difficulty scores with a one-way analysis of variance (ANOVA) (5 levels of task demand).

RESULTS

Cardiovascular Results

Cardiovascular baseline. We used a one-way ANOVA (five levels of difficulty) for any baseline index to examine whether the groups differed in their rest periods. No differences between groups were found (ps > .17). Means and standard errors of baseline scores by difficulty levels are reported in Table 1.

Cardiovascular reactivity. The quadratic contrast of difficulty levels on PEP reactivity scores was significant, $F(1, 105) = 7.69$, $p < .01$, $\eta_p^2 = .07$, and captured all significant variance (residual, $F < 1$). PEP reactivity scores were high for intermediate level ($M = –8.22$, $SE = 3.57$) and low for other levels of difficulty ($M = –1.56$, $SE = 1.71$ for very easy; $M = –0.72$, $SE = 0.86$ for easy; $M = –3.21$, $SE = 1.71$ for very difficult; $M = 2.36$, $SE = 0.95$ for impossible) (Figure 3). Decomposing this curvilinear effect using focused comparisons between conditions revealed a significant difference between easy and intermediate, $F(1, 42) = 4.26$, $p < .04$, $\eta_p^2 = .09$, intermediate and impossible, $F(1, 42) = 8.49$, $p < .001$, $\eta_p^2 = .17$, and very difficult and impossible, $F(1, 42) = 8.10$, $p < .01$, $\eta_p^2 = .16$, levels. No significant difference was found between other levels (ps > .09).

We examined whether PEP reactivity scores differed as a function of time-on-task with complementary ANOVAs. No difference was found (ps >.19).

Behavioral Results

Error rate. The linear contrast of difficulty levels on error rate was significant, $F(1, 105) = 936.76$, $p < .001$, $\eta_p^2 = .90$, (residual, $F < 1$). Means and standard errors were $M = 10.18$ (2.35) for very easy, $M = 22.69$ (2.90) for easy, $M = 52.03$ (3.29) for intermediate, $M = 80.94$ (2.31) for very difficult, and $M = 100$ (0) for impossible (Figure 4). Decomposing this effect by means of one-way ANOVAs showed significant differences across all conditions (ps < .01 and $\eta_p^2 > .21$).

MT. The quadratic contrast of difficulty levels on MT was significant, $F(1, 84) = 9.58$, $p < .01$, $\eta_p^2 = .09$ (residual, $F < 1$). Means and standard errors were $M = 1238.25$ (52.32) for very easy, $M = 1304.37$ (42.36) for easy, $M = 1322.44$ (48.59) for intermediate, and $M = 1112.30$ (32.72) for very difficult (Figure 4).

Complementary, focused comparisons between conditions revealed a significant difference of very difficult level with very easy, $F(1, 42) = 4.16$, $p < .05$, $\eta_p^2 = .09$, easy, $F(1, 42) = 12.87$, $p < .001$, $\eta_p^2 = .23$, and intermediate, $F(1, 42) = 12.87$, $p < .001$, $\eta_p^2 = .23$, levels. No other difference was significant (ps > .33).

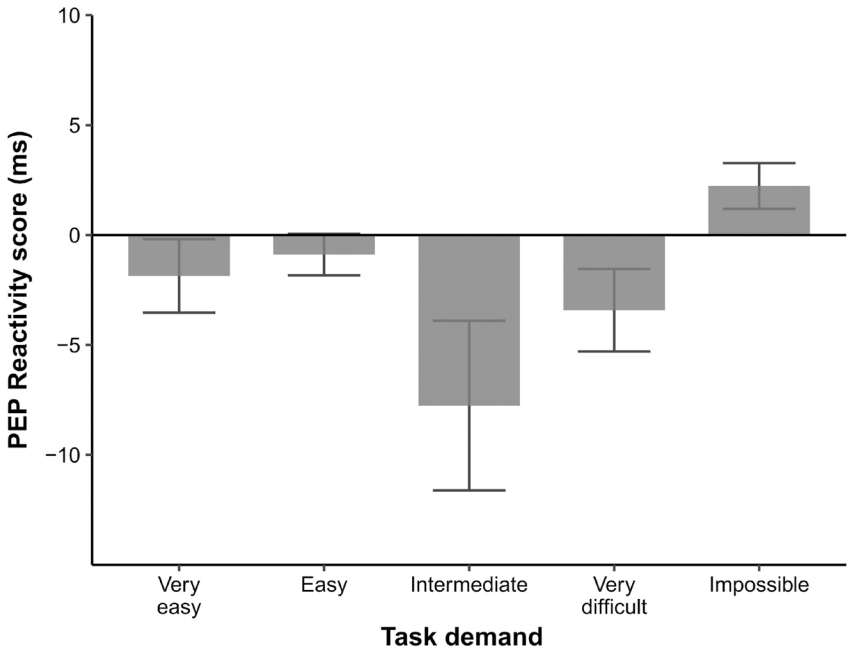


Figure 3. Mean pre-ejection reactivity scores (milliseconds) by task demand. Error bars represent standard error.

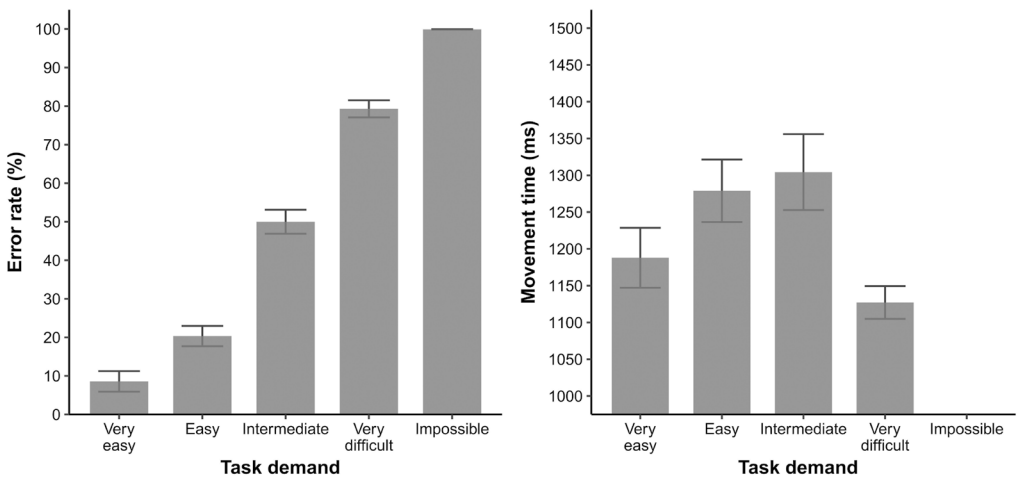


Figure 4. Means of error rate (percentage) and MT (milliseconds) by task demand. Error bars represent standard error.

Subjective Results

Perceived difficulty. The one-way ANOVA revealed a significant effect of task difficulty on the perceived difficulty, $F(4, 105) = 12.08$, $p < .001$, $\eta_p^2 = .10$. Perceived difficulty increased across different levels. Means and standard errors were $M = 2.04$ (0.10) for very easy, $M = 2.25$ (0.14) for easy, $M = 2.41$ (0.12) for

intermediate, $M = 2.49$ (0.13) for very difficult, and $M = 3.35$ (0.17) for impossible.

DISCUSSION

We investigated the curvilinear relationship of mental workload on behavioral and physiological adaptability, as postulated by Hancock and Szalma (2006) (Figure 1). To address this research topic, we used cardiac PEP reactivity, a more reliable and valid index of mental effort mobilization (Richter et al., 2008).

Manipulation of Mental Workload

The perceived difficulty scores and error rate increased across the five levels of task demand. These results suggest that experimental manipulation of mental workload was successful.

PEP Reactivity and Mental Workload

PEP reactivity supported our predictions about task demand effects derived from Hancock and Szalma's (2006) model. The curvilinear effect of mental workload on PEP reactivity was significant (Figure 3). PEP reactivity scores were high at the intermediate level and low at other levels of task demand. In contrast to PEP reactivity, task demand effects on HR, SBP, and DBP were not significant. The lack of effects of these cardiovascular measures is not surprising, as PEP is an index of beta-adrenergic influence on the heart. Moreover, it is important to note that decreases in PEP reflect increases in beta-adrenergic activity only when HR and blood pressure are stable or increase (Sherwood et al., 1990). As HR and blood pressure were stable in the present study, we concluded that beta-adrenergic effect caused the present PEP responses, reflecting mental effort mobilization (Wright, 1996).

Behavior, PEP Reactivity, and Mental Workload

As expected, we observed a linear effect of task difficulty on error rate and a curvilinear effect on MT (Figure 4). Moreover, PEP reactivity scores increased up to the intermediate level (Figure 3). Taken together, these results suggest that participants mobilized more mental effort to compensate for the decline in performance and to cope with the increase in task demand. However, when task demand increased above the intermediate level (i.e., very difficult level), error rate continued to increase but MT decreased. Moreover, the PEP reactivity scores decreased for the very difficult level, suggesting that the participants invested less mental effort. In fact, the speed–accuracy tradeoff observed for very difficult level as a function of an increase of MT could not be interpreted as a higher mobilization of mental effort but as a disengagement. In agreement with the model of Hancock and Szalma (2006), when task demand increased and approached overload, the participants reduced mental effort investment by using a behavioral strategy. Because the participants believed that they had a low probability to attain the target, they probably decided to change the strategy and move faster to the targets (Hoffmann, 2011; Hoffmann et al., 2013).

These results are consistent with previous studies on the role of behavioral adaptation to keep mental effort within acceptable limits. For instance, Loft, Sanderson, Neal, and Mooij (2007) reviewed all factors predicting mental workload of air traffic controllers to propose a model emphasizing the controllers' ability to maintain their mental workload under acceptable levels by means of different behavioral adaptation. Interestingly, they identified different means of behavioral adaptations in real tasks, such as the reordering of task priorities or the management of resources through explicit control of the airspace.

The Relationship Between Subjective Measures of Task Demand and Mental Effort

A common idea postulated by the MIT (Brehm & Self, 1989) is that mental effort mobilization is a linear function of subjective task demand, as long as success is possible and justified. This relation was confirmed in several studies using cardiovascular measures and PEP reactivity (Richter et al., 2016). However, in the present study, perceived difficulty increased across the level of task demand, and we observed an increase of PEP reactivity up to the intermediate level and a decrease after the intermediate level. Contrary to the expectations

and results of studies using the MIT theory, PEP reactivity related to effort did not increase up to the maximally possible and justified level and thus did not collapse afterward. Instead, the present study revealed a curvilinear relationship with a higher level of effort mobilization at an intermediate level of task difficulty, whereas the MIT predicted a shark-fin shaped function with a higher level of effort mobilization at a very difficult level.

This difference could stem from the nature of the mental workload used in the present experimental task. Based on Hoffmann (2011) and Hoffmann et al. (2013), we used shrinking targets until reaching one pixel. In fact, the time pressure was high in the very difficult condition, and the participants had faster MTs to capture the target prior to it disappearing. At this very difficult level, a speed–accuracy trade-off was observed; hence, it was not necessary for the participants to invest more mental effort. Such speed–accuracy trade-off was generally not present in the studies conducted with the MIT. For example, neither speed–accuracy trade-off nor change in behavioral adaptability across different levels of difficulty was found in the study of Richter et al. (2008). Therefore, the participants had probably used the same strategy across different levels of difficulty. They invested greater mental effort to reduce the decline in performance up to the maximally possible and justified level of effort. In professional and performance-critical environments, Hancock and Szalma (2006) found that goal distance and time pressure are two main sources of mental workload for the operators. The operator's adaptability at the behavioral and physiological level is predicted to counteract the effects of goal distance and time pressure. Further studies should test whether time pressure influences adaptability and the relationship between subjective difficulty and mental effort.

The second explanation for this curvilinear relation may be related to differences between difficulty levels in subjective importance. Subjective difficulty and subjective importance are generally moderately correlated (Eccles & Wigfield, 1995). When the participants score higher on subjective difficulty, they score lower on subjective success. Referring to MIT, subjective success determines the level of maximally justified effort. In the present study, one possibility is that participants have invested mental effort as a function of subjective difficulty up to the intermediate level. However, above this level, mobilization of effort may have been determined by the low level of maximally justified effort, resulting in a progressive disengagement and a curvilinear effect. Further studies should compare the subjective success and difficulty to disentangle these respective contributions.

ACKNOWLEDGMENT

This research was supported by grants from Région Occitanie and Université de Toulouse, INU Champollion.

KEY POINTS

- In ergonomics, Hancock and Szalma (2006) regularly postulated behavioral and physiological adaptability as a function of mental workload. However, these predictions never received a clear experimental validation.
- We tested Hancock and Szalma's (2006) predictions of mental workload model with a cardiovascular measure related to mental effort mobilization. We used the pre-ejection period (PEP), an index of beta-adrenergic sympathetic impact.
- The results provided the first evidence for a curvilinear effect of mental workload on PEP reactivity and behavioral adaptability as a function of mental workload.

REFERENCES

Benschop, R. J., Nieuwenhuis, E. E., Tromp, E. A., Godaert, G. L., Ballieux, R. E., & van Doornen, L. J. (1994). Effects of beta-adrenergic blockade on immunologic and cardiovascular changes induced by mental stress. *Circulation*, *89*, 762–769.

Berntson, G. G., Lozano, D. L., Chen, Y. J., & Cacioppo, J. T. (2004). Where to Q in PEP. *Psychophysiology*, *41*, 333–337.

Borghetti, B. J., Giametta, J. J., & Rusnock, C. F. (2017). Assessing continuous operator workload with a hybrid scaffolded neuroergonomic modeling approach. *Human Factors*, *59*, 134–146.

Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, *40*, 109–131.

Brookhuis, K. A., De Waard, D., & Fairclough, S. H. (2003). Criteria for driver impairment. *Ergonomics*, *46*, 433–445.

Capa, R. L., Audiffren, M., & Ragot, S. (2008). The effects of achievement motivation, task difficulty, and goal difficulty on physiological, behavioral, and subjective effort. *Psychophysiology*, *45*, 859–868.

Cegarra, J., & Chevalier, A. (2008). The use of Tholos software for combining measures of mental workload: Toward theoretical

and methodological improvements. *Behavior Research Methods*, *40*, 988–1000.

Cegarra, J., & Hoc, J. M. (2006). Cognitive styles as an explanation of experts' individual differences: A case study in computer-assisted troubleshooting diagnosis. *International Journal of Human-Computer Studies*, *64*, 123–136.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, *21*, 215–225.

Eubanks, L., Wright, R. A., & Williams, B. J. (2002). Reward influence on the heart: Cardiovascular response as a function of incentive value at five levels of task demand. *Motivation and Emotion*, *26*, 139–152.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *67*, 381–391.

Fitts, P. M., & Peterson, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology*, *67*, 103–112.

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Cognitive processes and performance* (Vol. 2, pp. 1–49). Oxford, UK: John Wiley.

Hancock, P. A. (2013). In search of vigilance: The problem of iatrogenically created psychological phenomenon. *American Psychologist*, *68*, 97–109.

Hancock, P. A., & Caird, J. K. (1993). Experimental evaluation of a model of mental workload. *Human Factors*, *35*, 413–429.

Hancock, P. A., & Chignell, M. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*, 647–658.

Hancock, P. A., & Szalma, J. L. (2006). Stress and neuroergonomics. In R. Parasuraman & M. Rizzo (Eds.), *The brain at work* (pp. 195–206). Oxford, UK: Oxford University Press.

Hancock, P. A., & Szalma, J. L. (2008). *Performance under stress*. Aldershot, UK: Ashgate Publishing.

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors*, *31*, 519–537.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology: Human mental workload* (Vol. 52, pp. 139–183). Oxford, UK: North-Holland.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, *45*, 73–93.

Hoffmann, E. R. (2011). Capture of shrinking targets. *Ergonomics*, *54*, 519–530.

Hoffmann, E. R., Chan, A. H., & Dizmen, C. (2013). Capture of shrinking targets with realistic shrink patterns. *Ergonomics*, *56*, 1766–1776.

Johnson, W. W., & Hart, S. G. (1987). Step tracking shrinking targets. *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 248–252). Santa Monica, CA: Human Factors and Ergonomics Society.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kelsey, R. M. (2012). Beta-adrenergic cardiovascular reactivity and adaptation to stress: The cardiac pre-ejection period as an index of effort. In R. A. Wright & G. H. E. Gendolla (Eds.), *How motivation affects cardiovascular response: Mechanisms and applications* (pp. 43–60). Washington, DC: American Psychological Association.

Levick, J. R. (2003). *An introduction to cardiovascular physiology*. London, England: Oxford University Press.

Llabre, M. M., Spitzer, S. B., Saab, P. G., Ironson, G. H., & Schneiderman, N. (1991). The reliability and specificity of delta versus residualized change as measures of cardiovascular reactivity to behavioral challenges. *Psychophysiology*, *28*, 701–711.

Loft, S., Sanderson, P., Neal, A., & Mooij, M. (2007). Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. *Human Factors*, *493*, 376–399.

Lozano, D. L., Norman, G., Knox, D., Wood, B. L., Miller, B. D., Emery, C. F., & Berntson, G. G. (2007). Where to B in dZ/dt. *Psychophysiology*, *44*, 113–119.

Marinescu, A. C., Sharples, S., Ritchie, A. C., Sánchez López, T., McDowell, M., & Morvan, H. P. (2018). Physiological parameter response to variation of mental workload. *Human Factors*, *60*, 31–56.

Matthews, J., Altman, D. G., Campbell, M. J., & Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, *300*, 230–235.

Moray, N., Dessouky, M. I., Kijowski, B. A., & Adapathya, R. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors*, *33*, 607–629.

Obrist, P. A. (1981). *Cardiovascular psychophysiology: A perspective*. New York, NY: Plenum Press.

Parasuraman, R., & Hancock, P. A. (2001). Adaptive control of mental workload. In P. A. Hancock & P. A. Desmond (Eds.), *Human factors in transportation: Stress, workload, and fatigue* (pp. 305–320). Mahwah, NJ: Lawrence Erlbaum.

Richter, M., Friedrich, A., & Gendolla, G. H. (2008). Task difficulty effects on cardiac activity. *Psychophysiology*, *45*, 869–875.

Richter, M., Gendolla, G. H. E., & Wright, R. A. (2016). Three decades of research on motivational intensity theory: What we have learned about effort and what we still don't know. In A. J. Elliot (Ed.), *Advances in motivation science: Advances in motivation science* (Vol. 3, pp. 149–186). San Diego, CA: Elsevier Academic Press.

Sherwood, A., Dolan, C. A., & Light, K. C. (1990). Hemodynamics of blood pressure responses during active and passive coping. *Psychophysiology*, 27, 656–668.

Victor, R. G., Seals, D. R., Mark, A. L., & Kempf, J. (1987). Differential control of heart rate and sympathetic nerve activity during dynamic exercise. Insight from intraneural recordings in humans. *The Journal of Clinical Investigation*, *79*, 508–516.

Wickens, C. D. (2001). Workload and situation awareness. In P. A. Hancock & P. A. Desmond (Eds.), *Stress, workload, and fatigue* (pp. 443–450). Mahwah, NJ: Lawrence Erlbaum.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*, 449–455.

Wright, R. A. (1996). Brehm's theory of motivation as a model of effort and cardiovascular response. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 424–453). New York, NY: Guilford Press.

Wright, R. A., & Kirby, L. D. (2001). Effort determination of cardiovascular response: An integrative analysis with applications in social psychology. *Advances in Experimental Social Psychology*, *33*, 255–307.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*, 1–17.

Charlotte Mallat is a doctoral student in the cognitive ergonomics PhD program at Federal University of Toulouse, INU Champollion, SCoTE Lab, Albi, France. She received an MS in ergonomics from the INU Champollion, France, in 2016.

Julien Cegarra is a full professor in the Department of Psychology at INU Champollion, France, and director of the Science of Cognition, Technology, and Ergonomics (SCoTE) Lab in Albi, France. He received his PhD in ergonomics from University of Paris 8 in France in 2004.

Christophe Calmettes is an assistant professor in Computer Engineering at INU Champollion, France, and member of Science of Cognition, Technology, and Ergonomics (SCoTE) Lab in Albi, France. He received his PhD in automatics from University Toulouse 3 in France in 1997.

Rémi L. Capa is an assistant professor in the Department of Psychology at INU Champollion, France, and member of Science of Cognition, Technology, and Ergonomics (SCoTE) Lab in Albi, France. He received his PhD in psychology on mental effort from University of Poitiers (CNRS) in France in 2007.

Date received: November 13, 2018
Date accepted: May 16, 2019

Concerns About Verbal Communication in the Operating Room: A Field Study

Ehsan Garosi, Tehran University of Medical Sciences, Iran,
Reza Kalantari, Shiraz University of Medical Sciences, Iran,
Ahmad Zanjirani Farahani, Arak University of Medical Sciences, Iran,
Mojgan Zuaktafi, Shiraz University of Medical Sciences, Iran,
Esmaeil Hosseinzadeh Roknabadi, Iran University of Medical Sciences, Tehran, Iran,
and **Ehsan Bakhshi**, Kermanshah University of Medical Sciences, Iran

Objective: To assess verbal communication patterns which could contribute to poor performance among surgical team members in an operating room.

Background: There exist certain challenges in communication in health care settings. Poor communication can have negative effects on the performance of a surgical team and patient safety. A communication pattern may be associated with poor performance when the process of sending and receiving information is interrupted or the content of conversation is not useful.

Method: This cross-sectional field study was conducted with 54 surgical teams working in two Iranian hospitals during 2015. Two observers recorded all verbal communications in an operating room. An in-depth assessment of various annotated transcripts by an expert panel was used to assess verbal communication patterns in the operating room.

Results: Verbal communication patterns which could contribute to poor performance were observed in 63% of the surgeries, categorized as communication failures (17 events), protests (23 events), and irrelevant conversations (164 events). The anesthesiologists and the circulating nurses had the most concerning communication patterns. The failure of devices and poor planning were important factors that contributed to concerning patterns.

Conclusion: Concerning patterns of verbal communication are not rare in operating rooms. Analyzing the annotated transcripts of surgeries can conduce to identifying all these patterns, and their causes. Concerning communication patterns can be reduced in the operating room by providing interventions, properly planning for surgeries, and fixing defective devices.

Application: The method used in this study can be followed to assess communication problems in operating rooms and to find solutions.

Keywords: communication analysis, team communication, patient safety, surgical care

Address correspondence to Reza Kalantari, PhD Student of Ergonomics, Department of Ergonomics, Faculty of Health, Shiraz University of Medical Sciences, Razi Boulevard, Shiraz, Iran; e-mail: rkalantari@sums.ac.ir.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 940–953
DOI: 10.1177/0018720819858274
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Clinical communication is a complex issue (Miller, Weinger, Buerhaus, & Dietrich, 2010). There are certain communicational challenges among health care professionals (such as lack of confidence, standardization, or experience; Foronda, MacWilliams, & McArthur, 2016). In health care settings, medical errors, delay in treatment processes, misdiagnosis, and adverse events such as patient injury and death can be enhanced by communication problems which can negatively affect performance measures such as time and accuracy (The Joint Commission, 2015; Interprofessional Education Collaborative [IPEC] Panel, 2011; Australian Commission on Safety and Quality in Health Care [ACSQHC], 2012). Based on the reports of The Joint Commission, communication flaws were observed in 64% of the root causes of sentinel events reported from 1995 to 2005 (Joint Commission Resources, 2008). Failure in communication among clinical teams can result in medical errors (Safran, Miller, & Beckman, 2006).

Communication constantly happens in and out of the operating room, as well as during the preoperative, intraoperative, and postoperative phases of care, hence involving very important issues such as patient status, surgical events, the plan of care, and so forth (Greenberg et al., 2007) and influencing team performance (Blom et al., 2007). Previous studies on problematic communication in the context of the operating room reported issues on low quality, inappropriate and ineffective communication, miscommunications, and communication breakdowns, signaling at least one defect in the transmission or content of information. Inappropriate communication is a

major contributing factor to errors in surgeries (Parker, Wadhera, Wiegmann, & Sundt, 2009; Wilson, Whyte, Gangadharan, & Kent, 2017). Ineffective communication in surgery was seen in 30% of the information exchanges, and approximately one third of these interactions put patient safety at risk (Lingard et al., 2004); it is known surgical teams which are not trained in nontechnical skills are associated with higher surgical mortality risk (Neily et al., 2010). Furthermore, communication delays can negatively affect the efficiency of the distributed team collaboration (Fischer & Mosier, 2014). The reasons for concerning verbal communication in the operating room are very complicated (Firth-Cozens, 2004), due to factors such as the variety of professional groups that work together and the different cultures such as a hierarchy like military settings (Lingard et al., 2004). Many of the communication breakdowns during the surgery tend to be verbal (Greenberg et al., 2007), so analyzing verbal communication in the operating rooms is necessary.

In a former classification of problematic communication types in the operating room, Lingard et al. categorized four communication failure types, namely occasion (wrong timing), content (missing or inaccurate information), purpose (unresolved issues), and audience (missing key individuals). They suggested communication failures occur when there is a flaw in the abovementioned rhetorical factors (Lingard et al., 2004). In this study, we assumed there are more problematic communication issues which may be associated with poor performance. For example, according to the former taxonomy (Lingard et al., 2004), an irrelevant communication may not be considered as "failure," although it may put surgical performance at risk as it is known that irrelevant conversations can lead to poor performance and the occurrence of unwanted events in the operating room (Persoon, Broos, Witjes, Hendrikx, & Scherpbier, 2011). Although speech acts can be helpful to share situation-related information in the operating room (Parush et al., 2011) and harmonize team members, such acts should be used carefully as the aseptic technique recommends team members to avoid talking during surgery to prevent possible infections (Ahsan, 1997) because verbal conversation can increase the risk of infection in the operating room by dispersing more oral bacteria. Much communication does not always lead to better performance (Jentsch, Salas, Sellin-Wolters, & Bowers, 1995), and in some teams, the quality of communication can be more important than its quantity (Barnlund, 1959). It is known that distractions and errors are related to problematic communication (Halverson et al., 2011). Accordingly, communication patterns which could be contributed to poor performance in the operating room are not limited to failures and have a wider range which can be investigated with adopting a more extensive view.

This study was conducted to assess verbal communication patterns which could be contributed to poor performance among surgical team members in operating rooms. Based on the literature and theories of communication, a communication pattern is assumed to be "concerning" or "associated with poor performance" when the process of sending and receiving information is interrupted or when the content is not useful and may put performance measures such as speed, accuracy, and satisfaction (related to surgical team performance) at risk. The reasons and different types of concerning patterns of communication were analyzed using field observation and annotated transcripts for each individual surgery, with no need to recall the surgery process and events for data extraction. We attempted to group and measure the prevalence of the verbal communication patterns which may be associated with poor performance, find the reason for the concerning patterns, and identify the involved team members.

METHOD

Study Design

This cross-sectional and observational study was conducted in the orthopedic surgical wards of two Iranian hospitals during the summer of 2015. Two trained observers recorded all communication events in the operating room. Then, the annotated transcript of the process of each surgery was written. An expert panel consisting of five specialists in the domains of human factors, surgery, and psychology was used to extract and analyze all communication patterns. In the next step, the communication

patterns were grouped, and the frequencies of the patterns which could be contributed to poor performance were surveyed based on communication theories and surgery guidelines.

Participants

The studied surgical wards were selected randomly. Only one surgery type (orthopedic) was selected to make the results more comparable. Ten practice surgeries and after that 54 of 60 possible surgical teams (including 114 participants) were selected using the cluster random sampling method. The clusters included the surgeries of knee, shoulder, hand, fibula, thigh, and hip. Each of the studied teams differed from each other by at least one member, and 50 of 54 teams (92.6%) had overlapping members. The inclusion criteria were the willingness of all team members and patients to participate, and that each team was composed of six practitioners from surgery, anesthesiology, and nursing domains. Surgeon, surgeon assistant, anesthesiology specialist, anesthesiology technician (anesthesiologist), scrub nurse, and circulating nurse were the team members. The exclusion criterion was the presence of trainees in the operating room since the hospitals were not teaching centers. Six teams were excluded, two due to the unwillingness of the surgeons, and four because of the trainees.

Surgery Process

Generally, in surgeries, the patient is transferred to the operating room where the anesthesiologist and circulating nurse are present. The circulating nurse checks the patient's identity and surgery location, whereas the anesthesiologist asks questions about the patient's last meal and drug allergies. Then with the coordination of the anesthesiology specialist, the anesthesia program is determined and applied. Usually, the surgeon checks the patient's status before surgery. The operating room nurses prepare surgery location. The surgeon and the surgery assistant work on the surgery spot in the sterile area, whereas the scrub nurse helps them through providing the essential equipment or anything necessary. During the surgery, the anesthesiologist monitors the patient's consciousness and controls his or her airway, whereas the anesthesiology specialist sometimes checks the overall condition of the patient. The circulating nurse controls all events happening outside the sterile area and provides team members with anything necessary. Following the surgery, the patient is transferred to the recovery room.

Data Collection

Data were collected through observing and recording surgery events (the sentence method of note-taking and voice recording) in the operating room. All the conversations and communication events in the operating room were recorded from the beginning to the end of the surgeries. Two observers, including a specialist in human factors and a technician of the operating room who were experienced and trained to record data in this context, conducted this step. They entered the operating rooms before the surgical team members and exited after them. They dressed in operating theater uniforms and stood near the surgery bed (but in the unsterile area), so they could observe better all the surgical events and communications. They stood where they did not interrupt the normal activities and distract the team members. Ten practice surgeries were selected and observed before collecting the data during the 54 main surgeries, so that the observers could practice recording events without missing data; it further helped observe the same population and coordinated the two observers. No data were extracted from the 10 surgeries. Only verbal communications were studied because the assessment of nonverbal ones required specific settings difficult to provide and was unacceptable for hospital managers. However, to avoid misunderstanding of failures, any relevant and effective information from the context and environment such as situations, devices, task requirements, and nonverbal responses was recorded.

The Expert Panel

To assess the annotated transcripts, we organized an expert panel using group processes including five experts: two experts in psychology with more than 10 years of training, two specialists in human factors with 5 years of research experience in communication and teamwork in the operating room, and a consultant surgeon with 18 years of work experience. Prior to the start of the group processes, the members were informed with a package of information on all of the models of communication which some

of them were specific to interpersonal communication, communication in the operating room, surgical procedure, orthopedic surgeries, work standards in the operating room, rules of each participant, and safe surgery. The members were asked to read the package prior to data interpretation. During this time, the transcripts containing all communications, and the related contextual events for each surgery were prepared.

Communication Models

Based on the decision of the expert panel, the transactional model of communication and the information theory were selected to extract and analyze the communication patterns. According to the transactional model (Barnlund, Akin, Goldberg, & Myers, 1970, pp. 83–92) which is applicable in communication between two or more individuals, each person is both a speaker and a listener, further reacting based on factors such as attitudes, cultural beliefs, and experience. A sender and a receiver of information are interdependent and necessary to continue communication, so there is a need for feedback which can also be a nonverbal response.

The information theory of communication is a comprehensive model which refers to technical (accuracy in transference of the information), semantic (interpretation of the meaning by the receiver), and effectiveness (leading the message to the desired goal) problems in the communication process. Based on this model, communication is the transmission of information from a sender to a receiver. First, the information source selects a message; the transmitter then changes the message to a signal conveyed by a communication channel; ultimately, the receiver changes the transmitted signal back to the message and hands the message on to the destination. For example, when one person talks to another, his or her brain is the information source, his or her vocal system is the transmitter; the brain of the interlocutor is the destination, and his or her ear is the receiver (Shannon & Weaver, 1949, pp. 4–7).

Preparation of Data and the Structure of Patterns

To prepare the data, the human factors expert and the psychologist wrote the annotated transcripts for each surgery. None of the transcripts were excluded. The "narrow transcription approach" (which captures conversational interactions, diction, loudness, and talk overlap) was selected, and then literal transcripts were provided. Next, the communication patterns were extracted from the transcripts according to the sequence of the events.

The communication patterns were extracted over many sessions, where an in-depth assessment of the events was used to avoid possible biases. Every communication effort was considered as an event or speech act, such as request, announcement, question, reply, confirmation, and read-back (Parush et al., 2011). A simple sending, receiving, and possible feedback (each considered solely as an event) formed the shortest possible pattern considered as an uninterrupted and meaningful sequence of verbal interactions produced by at least two participants. A pattern is determined based on the start and end of an exchange surrounding a specific event or topic, which might include a request for an item. The length of each pattern was determined based on the content, communication process, and the sequence of events. Table 1 shows an example of a pattern, in a situation where the surgeon needs a Vicryl.

As we see in the table, the surgeon starts the communication process with a request, as a sender. The circulating nurse receives the information and provides feedback by responding to the surgeon. The information exchange continues until the specific topic and the pattern are both finished. A new sending of information about another topic or event starts a new pattern.

Data Analysis, Coding Scheme, and Group Discussion Process

At the beginning of the first session of the expert panel, the human factors specialist welcomed the group members and asked them to introduce themselves. The aim, topic, and rules were then described. All recorded communication events were assessed over seven sessions, each lasting 3 to 4 hr, where panel members discussed each event separately. The theoretical framework and the grounded theory approach principles were used in the process of data gathering and analysis; therefore, prior to any

TABLE 1: An Example of a Communication Pattern Between a Surgeon and a Circulating Nurse

Surgeon: "Please give me a Vicryl, which sizes are available?"
Circulating nurse: "1 and 2; I think."
Surgeon: "Give me 1, it is good."
Circulating nurse: "Here it is."
Surgeon: "Thank you."
Circulating nurse: "You're welcome."

analysis, the data were broken into small pieces, and the categories emerged purely from the data which could be divided into distinguishing properties. Furthermore, an open-coding process created the concepts, and the iterative analysis of data was conducted, where data came first.

The recorded communication events were coded into two categories by the expert panel: (A) patterns not associated with poor performance and (B) patterns which might be associated with poor performance. This coding scheme is derived from three resources: (1) transactional theory of communication which points problems in sending, receiving, or feedback; (2) information theory of communication which highlights technical, semantic, and effectiveness problems, and (3) standards, rules, norms of working instructions in the operating room, guidelines for safe surgery, and the limitations of communication in operating rooms. These three resources generated certain questions conducive to data extraction, such as "Is the information transfer happening accurately? Is the desired meaning transferred? Is the message effective for the receiver? Is the message necessary at all? Can the message distract team members?" The pattern with at least one problem regarding the employed resources was coded as "possibly associated with poor performance."

The transcripts of the first 10 out of the 54 main surgeries were coded to ensure coding consistency and to coordinate the coders. Interrater reliability among the coders was calculated, and it was at an acceptable level as the absolute agreement between coders was 90%. The remaining 44 surgery transcripts were then coded under the supervision of one researcher. In the first coding process, the expert panel extracted the concerning patterns. After that, all patterns were checked again and discussed by the panel members, and final changes were made. More consistency in the coding process was obtained from the discussion among coders. The patterns possibly contributed with poor performance were listed, discussed, and deeply analyzed during group discussions, and ultimately coded. The patterns were divided based on the mentioned models of communication, the cause of the occurrence, similarity, and difference. The concerning patterns were coded as (B1) sending information is problematic, (B2) receiving information is problematic, (B3) conversations are not related to the patient or the operating room, and (B4) conservations which could be avoided due to operating room limitations in communication.

Group decision making and coding processes were conducted by 1 to 5 scoring method for each communication pattern from "strongly disagree" to "strongly agree" regarding inappropriateness of them. Opinions and preferences of the expert panel members about communication patterns were summed up and averaged to reach a decision. The patterns with mean scores of four or more were categorized as concerning. Different ideas were then evaluated through group discussions. The necessity and effectiveness of conversations regarding operating room situation were led by the surgeon, whereas the behavior of team members and quality of communications were recorded by psychologists and human factors experts based on their special viewpoint of behavioral sciences. The results of each session were recorded by a secretary. After the analysis, all results were checked and discussed, and then the prevalence of concerning patterns was determined. The explanation of qualitative research steps is provided in Table 2.

The study was approved by the ethics committee of Tehran University of Medical Sciences. The observers entered the surgical wards with the permission of the hospital managers. The researchers formally introduced themselves to the surgical teams and asked for permission for data gathering. All participants and patients signed the informed consent form before the data collection (before entering the operating room). The observers did not disturb the surgical team members during the surgery through speech or actions.

TABLE 2: The Process of the Qualitative Analysis

Steps	Actions
1	1 month preparation using an information package about communication models, related literature, and the guidelines and working condition of the operating room.
2	Extracting communication patterns from the literal transcripts based on communication models.
3	Defining the coding scheme based on communication models, related literature, the guidelines of surgery, and working condition of the operating room as (A) may be associated with poor performance and (B) not associated with poor performance.
4	Analysis of the first 10 surgeries (10 of 54) to ensure coding consistency, coordination of the coders, and reliability assessment.
5	Analysis of all 54 surgeries to extract the communication patterns which may be associated with poor performance.
6	Dividing the patterns which could be contributed to poor performance based on the mentioned models of communication, cause of the occurrence, similarity, and the difference in the concerning patterns.

RESULTS

The time length average and standard deviation of the 54 surgeries was 93 ± 27 min. All surgeries were successful. The total number of communication patterns was 624; 204 (32.69%) of which were detected as possibly associated with poor performance. Thirty-four surgeries (63%) contained at least one concerning communication pattern.

After an in-depth assessment of all the literal transcripts, the expert panel divided the concerning communications into three groups: communication failures, protests, and irrelevant conversations. Communication failures consisted of the communications in which there was a problem in sending or receiving information or when there was no feedback (based on transactional communication model). Feedback refers to a verbal response or any act to a request or command. A protest is any statement or action expressing disapproval of or objection to something, such as the performance of a surgical team member or the defects of a device. Irrelevant conversations were communications on topics other than the surgery or the patient.

In this study, there were 17 events of communication failure and 23 events of protest that occurred during the 54 surgical operations. Furthermore, in 28 surgeries, conversations irrelevant to the surgery or the patient were observed (164 events). All the recorded communication events belonged to these three groups. At least one category of concerning communication was found in 42 surgeries (78%), and two or three categories were identified in 16 surgeries (30%).

Communication Failures

Communication failures were divided into four categories according to their reason: (1) communication failure because verbal communication was so quiet that the addressed person (requester or responder) could not hear the message, (2) the communication failure occurred due to the lack of response because the addressed practitioner was absent, (3) communication failure was because of inattention to the team and the surgical process, and (4) communication failed due to the lack of response. In the third and fourth categories, the reasons for nonresponse were determined based on asking questions from team members following the surgery. In one pattern, everyone thought that another person would respond (bystander effect), which was measured by verbal confirmation of team members after the surgery and context information. These categories represent the lack of necessary feedback, so the communication loop remains incomplete based on the transactional model. As mentioned in the method section, all of the related information was recorded during the surgery while communicating. This information was conducive

TABLE 3: Communication Failure Types Based on Reason, Frequency, the Agent Involved, With an Example

Type of Communication Failure	Frequency (Times)	Agent (Times)	Example
1. Too quiet communication	1	Scrub nurse (1)	The scrub nurse received a request from the circulating nurse: "Would you please give me the Vicryl?" "Give what?" (Very quiet voice, not hearable)
2. No response because of absence	12	Anesthesiologist (8), circulating nurse (4)	When the patient felt pain, the surgeon asked the anesthesiologist to inject tranquilizer: "Please inject some tranquilizers." There was no response because the anesthesiologist was absent.
3. No response because of inattention	3	Circulating nurse (2), anesthesiologist (1)	The surgeon needed some tools, but the circulating nurse did not respond: "Give us some gauze pads." (Nothing) There was no response because of inattention by the circulating nurse.
4. No response because team members thought that another member would respond	1	All team members except the circulating nurse (1)	A surgical set must be brought by the circulating nurse. As he required some information about the surgery, he asked a question, but no one responded: "Which set do you want me to bring?" (Nothing) There was no response because all team members were waiting for others to answer.
Total	17	Anesthesiologist (9), circulating nurse (6), scrub nurse (1), all team members except the circulating nurse (1)	

to differentiating between situations in which a team member acts in response to a request, but says nothing. Every positive action regarding the request is considered as feedback; for example, when a surgeon requests a gauze pad and the circulating nurse provides the gauze pad without talking, the pattern ends and is not considered as problematic.

Communication failures were found in 30% (16) of surgeries. The examination of the communication failure agents revealed that 88% of these events were related to the inattention or absence of circulating nurse or anesthesiologist, whereas the communication directed toward different team members was approximately the same. In some observed surgeries, the circulating nurses or anesthesiologists did not pay attention to the surgical procedure; at certain points, they left the operating room without any notice or they were absent when they were needed. During some surgeries, they used cell phones or sat around and said nothing. Table 3 shows communication failure types based on reason, frequency, the agent involved, and an example of

TABLE 4: Details of Protests in the Operating Room, Frequency, Agent, and the Objector Involved

Objector (Times)	Agent (Times)	Frequency	Subject of Protest
1. Performance of another, the messiness of the operating room, slow and poor performances, leaving the operating room, absence during the handling of the patient, inattention, and arriving late	14	Circulating nurse (7) Anesthesiologist (3) All other team members (4)	Surgeon (6) Surgeon assistant (4) Anesthesiologist (2) Scrub nurse (1) Circulating nurse (1)
2. Devices that did not work correctly	4	The operating room devices	Surgeon (2) Scrub nurse (1) Surgeon assistant (1)
3. Annoying noise	1	No one inside the operating room	Scrub nurse (1)
4. Poor planning and the absence of the supervisor	4	No one inside the operating room	Surgeon (4)
Total	23	Circulating nurse (7) Anesthesiologist (3) All other team members (4) No one inside the operating room (9)	Surgeon (12) Surgeon assistant (5) Anesthesiologist (2) Scrub nurse (3) Circulating nurse (1)

each category. The agent refers to the inferred cause of communication failure.

Protests and Expressions of Discomfort

In this study, 23 events associated with protest and expressions of discomfort were recorded, divided into four categories according to the subject of the protest (Table 4). (1) Protest about the performance of others, occurring when one or more team members did not perform their duties correctly. (2) Protest because of the defects in the devices in the operating room, some of which were recorded and observed. Defective devices were common in the operating rooms, which distressed the practitioners. (3) Protest about an annoying noise which occurred during one surgery; however, the source of the noise was not in the operating room, but it distracted the practitioners anyway. (4) Protest arose from bad planning and a lack of supervision. The lack of clear planning and timing for the surgeries in the surgical wards bothered the team members. However, this type of protest was recorded only three times in all 54 surgeries. Furthermore, there was only one event of protest about the lack of supervision in the operating room. Some examples of each category are presented in Table 5, where the results indicate that the protests have emerged from both people (14 times) and general environment or system design (9 times).

Irrelevant Conversations

In all, 164 irrelevant conversations (about different issues) were reported in 28 (52%) surgeries; during 13 of these surgeries, irrelevant conversations happened when the surgeon was not present in the operating room. There were 10 surgeries that involved speaking about other patients; in eight of these surgeries, speaking on a cell phone; in two surgeries, playing music and singing; and in five surgeries, speaking to a person outside the operating room were observed and recorded. Some of the irrelevant conversations delayed team activities. For example, in one surgery, all the team members were waiting for the surgeon who was speaking with two other surgeons outside the operating room about irrelevant and personal issues. In some surgeries, talking about other patients

TABLE 5: Examples of Four Types of Protest

Type of Protests	Example
Performance of others	• The circulating nurse left the operating room without notice. When he came back, the surgeon assistant protested and said, "Tell me, why you are inattentive to the principles of work?" • The circulating nurse distracted the scrub nurse and talked about irrelevant issues. The surgeon protested and said, "Tell me, why you are disturbing her? Be quiet." • During the surgery, the anesthesiologist told the team, "The patient is regaining consciousness," and he asked them to finish the operation as quickly as possible. The surgeon protested and said, "What? Why did you not anesthetize him correctly? What are you doing?" • The surgeon required some tools and asked the circulating nurse to provide them. The circulating nurse did not respond quickly, and the surgeon protested and said, "You're doing it slowly. Come on. We don't have much time."
Defects in devices	• The surgeon was upset about the monitors used for arthroscopic surgery. "Oh! This monitor has been broken for several weeks. When do they want to fix this? It is really annoying."
Annoying noise	• There was a noisy discussion taking place outside the operating room that distracted the scrub nurse and caused him to make a mistake. The surgeon protested and said, "Guys, we are doing surgery and here is an operating room. Do not brawl here."
Poor planning, absence of the supervisor	• During surgery, someone came into the operating room, and said, "The next patient is ready for surgery and the (other) team is waiting for the surgeon." The surgeon protested and said, "Why did you start? Who let you? You shouldn't . . . I can't come . . . I am going to be busy for at least 20 minutes!" • The surgeon was ready and waiting for the scrub and circulating nurses. However, after 10 min, they were still not in the operating room. After they finally arrived, they did not perform as the surgeon expected. The surgeon summoned the supervisor of the ward and believed that all the bad situations were because of his absence. The surgeon protested and said, "What is going on here? In my whole career, I have never seen such a situation. The supervisor doesn't do anything here."

annoyed the team members. As well, the operation was delayed when the surgeon was speaking on a cell phone.

Nonconcerning Communication

A total of 420 communication patterns (67.31% of all patterns) were categorized as nonconcerning. These patterns were determined according to the three resources mentioned in the method section. Repeating the commands and requests of other team members, providing explanations about the surgical process, informing other team members at the change of surgery status, and providing feedbacks and confirmations are among the examples of nonconcerning communication.

DISCUSSION

This field study aimed to address concerning verbal communications among surgical team members. The analysis of the annotated transcripts showed that verbal communication patterns possibly contributed to poor performance are not rare in the operating room and can be classified as communication failures, protests, and irrelevant conversations. The resulted categories can be justified based on the information theory of Shannon and Weaver (1949, pp. 4–7). Based

on this theory, problems can occur in the accuracy of transmission, the precision of meaning that is conveyed, and the effectiveness of the message on the behavior. The latter is related to protests and irrelevant communication, whereas the first and second types of potential problems can occur in communication failures. Moreover, the transactional theory of communication justifies the problems in sending, receiving, and providing feedback in communication processes, hence covering the communication failure problems (Barnlund et al., 1970, pp. 83–92).

Former studies considered the communication failure types, whereas in this study, concerning communications and their causes were considered. Comprehensive consideration of communication and the causes of alarming patterns can be conducive to identifying defects and making suggestions about possible solutions; improving communication effectiveness in health care is regarded as a global priority (The Joint Commission, 2015; IPEC Panel, 2011; ACSQHC, 2012).

The prevalence of communication failures found in this study (32.7%) was higher than the study by Lingard et al. (2004), in which 30.6% of all communications had flaws. The difference can be due to adopting a more extensive view in this study. Other studies reported communication breakdowns and information loss in 62% (57% verbal communication) and 100% of the observed surgeries, respectively (Christian et al., 2006; Greenberg et al., 2007). All of the mentioned studies concluded that communication problems frequently take place in operating rooms, and it is important to determine the roots of the concerning patterns, maybe because of the complex work conditions of the operating rooms. The difference between the results may be due to the extent of the theories used to classify the communication events and the different cultural and demographic characteristics of the studied populations.

Examination of the agents involved in communication failures revealed the role of circulating nurses and anesthesiologists. A similar result was reported by Gardezi et al. (2009), who found that circulating nurses or anesthesiologists had less understanding of effective communication compared with other surgical team members. This result could be due to the workstations of the circulating nurses and anesthesiologists (outside the sterile area) and their unwillingness to actively participate in the surgery process and team communications. In some surgeries, they did not pay attention to other team members and caused some protests and delays. Using cell phones and speaking with people outside the operating room were actions which caused concerning communication patterns, although it is known that telephones, pagers, music, and other people can distract team members and cause undesirable events in the operating room (Persoon et al., 2011). Although failures in communication are also a part of an extensive system of relations, processes, and general organization (Lingard et al., 2004), we further considered communication as a key nontechnical skill, of which all team members should have sufficient mastery.

Protests and expressions of discomfort are avoidable communication patterns. As already mentioned, communication should be kept minimum due to the risk of infection. Protests originating from bad planning of surgeries and defects in devices are more avoidable than those due to the performance of others. Protests are not all negative and can even be useful in signaling the defects of a system and maintaining the culture of safety; however, the problem is that protests can distract team members and are possibly contributed to poor performance as shown by the analysis of annotated transcripts, where irrelevant communications were increased following the protests which can distract team members. Wheelock et al. (2015) suggested irrelevant conversations as main reasons for distractions. Some of the recorded protests were about defective equipment, known as common subjects of communications (Halverson et al., 2011). It is obvious that system design can influence communications as 39% of the protests in this study were caused by the components of the environment and system design, hence the importance of evaluating the tools that are part of a system. A broken device, or a surgery plan which is not properly arranged, does not need to be discussed during surgery, because it would not solve the problem at the moment and only distract team members. It is better that these issues be considered and reported before or after

the surgery. Expression of discomfort and protests might be considered important signals relevant to job design and safety; however, they can be avoided by a proper system design to reduce the risk of infection and distraction of other team members.

Irrelevant conversations were the most common concerning communication patterns observed in approximately half of the studied surgeries. The high prevalence of irrelevant conversations in the results could be because the design of this study recorded all possible events and neglecting surgical guidelines by team members. Irrelevant conversations may entail poor performances and the occurrence of unwanted events in the operating room (Persoon et al., 2011). In a controlled laboratory study, it was reported that irrelevant conversations were kept at a minimum to reach a better performance (Webster & Cao, 2006). One may argue that conversations can build rapport among team members. In the operating room, irrelevant conversations can distract team members and increase the probability of human errors. There are many sources of distraction in an operating room (Seelandt et al., 2014), which should be managed. Our observations showed that most of these conversations had a negative rather than a positive impact on many surgeries, ultimately distracting team members. Furthermore, trying to avoid talking during surgery is one of the most important parts of aseptic technique recommended to prevent transfer on microorganisms to the patient (Ahsan, 1997, p. 87), especially in orthopedic surgeries. Thus, it is necessary to filter avoidable verbal conversations and keep only positive ones.

Assessment of communicational behaviors revealed that some surgeons did not let other team members deviate the conversations from patient and surgery. Usually, surgeons are the surgical teams' formal leaders and take the lead in the content of the conversations and communications in an operating room. In one study, 87% of the operating room practitioners suggested that the surgeon was responsible for creating a silent environment to reduce distractions (Persoon et al., 2011). They have an important role in the quality and quantity of the exchanged information during surgery and can control and optimize the content of the communications to reduce distractions.

Similar to this research, most studies on operating room communication issues conducted in the United States, the United Kingdom, and Canada have shown that approximately one third of communications in operating rooms are problematic (Lingard et al., 2004; Parker et al., 2009). However, the findings of this study indicated that a vast number of concerning communications are irrelevant conversations, whereas the results of the former studies showed more communication failures and breakdowns. The numerous patterns of irrelevant conversations are possibly due to the limited attention to improving nontechnical skills such as communication with operating room staff, and the limited policies and supervisions on operating rooms in Iran (Kalantari et al., 2016). Such differences could also arise from cultural varieties.

The analysis of verbal communications is not easy because communication is not standardized in complex environments, such as operating rooms (Blom et al., 2007). For this reason, it is necessary to use reliable methods (Nagpal et al., 2010). An observational field study was used to record all the events from the real environment during the surgery process. This study design can determine who or what was the cause of the concerning verbal communications, which is conducive to deciding the necessary type of intervention. There was no missing information because the data collection method was not related to special items. Furthermore, using the same pair of observers for all surgeries increased the reliability of data collection, and the well-trained observers alongside clear rules in the recording procedure helped avoid bias. Using competent and highly trained participants from different domains in the expert panel helped increase validity. Analyzing the first 10 out of the 54 surgeries enabled the expert panel to recognize patterns and increase reliability. Using multiple coders and reviewing the results helped prevent bias. The presented categories, linked to surgical team members' concentration, are known as distracters which could cause adverse events. Moreover, representativeness (selection of teams using

random sampling), transferability (due to the comprehensive description of settings and processes), and using mixed method research have rendered the results generalizable to advanced theory and practice, although there is the need for further research.

The findings of this study can be useful in determining the reasons and situations that cause communication failures, complaints, and irrelevant conversations in operating rooms. Identifying the reasons for problems in communication can lead to the development of proper solutions or interventions. For example, educational interventions can be very effective to improve the communicational skills of operating room personnel (Kertesz, Walker, & Maliwat-Bandigan, 2018). Team members should be trained to provide suggestions, send and receive feedback, reflect feelings, and have a communication strategy (Wang, Chellali, & Cao, 2016). Feedbacks and nice comments (Webster & Cao, 2006), practicing the substantial properties of communication (Keyton & Beck, 2010), using structured communication protocols (Schiff, Moulder, Louie, & Toubia, 2016), closed-loop communication (Bowers, Jentsch, Salas, & Braun, 1998), and fixing defective devices in operating rooms (Halverson et al., 2011) can decrease communication flaws in the operating room. More generally, adopting rules and policies such as proper planning and timing can improve nontechnical skills such as communication in the operating room (Kalantari et al., 2016; Kalantari, Zanjirani Farahani, Garosi, Badeli, & Jamali, 2019).

LIMITATIONS

As the selected hospitals were not teaching centers, the presence of trainees in the operating room was not expected during surgery. Thus, their presence deviated communications in some surgeries, hindering the data collection process. For example, although the team members asked questions and instructed the trainees, they sometimes entered the operating room during the surgery process. Therefore, the results of this study are more generalizable to nonteaching hospitals because the surgeon assistants were experienced operation room technicians, not surgical residents. The preparation of annotated transcripts was a time-consuming process which is normal in such types of studies. Moreover, in two surgeries, the surgeons did not allow us into the operating room for personal reasons. And there was a selection bias because all teams were selected based on participants' willingness to participate in the study. We did not measure the impact of concerning communications on the patients.

CONCLUSION

The assessment of concerning verbal communications is a complex issue in the operating room. A field study based on the surgery process to assess verbal communication can identify the strengths and weaknesses of information exchange in an operating room and determine the causes and roots of concerning patterns. Verbal communications which could be contributed to poor performance are not rare in the operating rooms and can be categorized into communication failures, protests, and irrelevant conversations. Lack of attention to the surgery process, broken devices, annoying noise, poor planning, using cell phones in the operating room, and poor nontechnical skills are among the main causes of alarming communication patterns. Therefore, educational programs, providing functional devices, and setting policies can improve communication in an operating room and reduce errors. Future research can use this study design to assess the impact of different interventions and determine their effectiveness.

ACKNOWLEDGMENTS

We would like to express our sincere thanks to hospital officials and all operating room personnel who helped us with this research. We are also grateful to Ms. A. Kalantari for improving the English in the manuscript. This study was financially supported by Tehran University of Medical Sciences as part of a master's thesis. Ehsan Garosi is also affiliated with Iran University of Medical Sciences, Tehran, Iran. Reza Kalantari is also affiliated with Universitat Jaume I, Castellon, Spain.

APPENDIX

The Coding Scheme Used in the Study and the Final Categorization

1. Categorizing communication patterns using three resources:
    1. Information theory of communication
    2. Transactional theory of communication
    3. Standards, rules, norms of the operating room, guidelines of safe surgery, and the limitations of communication in the operating room
2. Categorizing communication patterns which could be contributed to poor performance based on:
    1. The mentioned theories of communication
    2. Cause
    3. Similarity
    4. Difference
3. Final categorizing of the problematic communications and prevalence study
    1. Communication failures
        a. Very quiet communication events
        b. No response due to the absence of the receiver
        c. No response due to inattention
        d. No response due to the bystander effect
    2. Protests
        a. Poor performance of others
        b. Defects in the devices
        c. Annoying noise
        d. Poor planning and absence of the supervisor
    3. Irrelevant conversations (any conversation not related to the patient and surgery).

KEY POINTS

- Communication plays an important role in patient safety in the operating room.
- Communication problems are not rare during surgical procedures (observed in 32% of all communications).
- Communication patterns possibly contributed to poor performance were categorized as communication failures, protests, and irrelevant conversations.
- In the observed surgeries, the behavior of the anesthesiologists and the circulating nurses contributed most to problems with communication.
- The defectiveness of devices or equipment in the operating room was a factor that caused concerning communication patterns.

REFERENCES

Ahsan, I. (1997). *Textbook of surgery*. Boca Raton, FL: CRC Press.

Australian Commission on Safety and Quality in Health Care. (2012). *Safety and quality improvement guide standard 6: Clinical handover*. Sydney, Australia: Author.

Barnlund, D. C. (1959). A comparative study of individual, majority, and group judgment. *The Journal of Abnormal and Social Psychology*, *58*, 55–60.

Barnlund, D. C., Akin, J., Goldberg, A., & Myers, G. (1970). *Language behavior: A book of readings in communication*. The Hague, The Netherlands: Mouton.

Blom, E., Verdaasdonk, E., Stassen, L., Stassen, H., Wieringa, P., & Dankelman, J. (2007). Analysis of verbal communication during teaching in the operating room and the potentials for surgical training. *Surgical Endoscopy*, *21*, 1560–1566.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors*, *40*, 672–679.

Christian, C. K., Gustafson, M. L., Roth, E. M., Sheridan, T. B., Gandhi, T. K., Dwyer, K., . . . Dierks, M. M. (2006). A prospective study of patient safety in the operating room. *Surgery*, *139*, 159–173.

Firth-Cozens, J. (2004). Why communication fails in the operating room. *BMJ Quality & Safety*, *13*(5), Article 327.

Fischer, U., & Mosier, K. (2014). The impact of communication delay and medium on team performance and communication in distributed teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, No. 1, pp. 115–119). Los Angeles, CA: SAGE.

Foronda, C., MacWilliams, B., & McArthur, E. (2016). Interprofessional communication in healthcare: An integrative review. *Nurse Education in Practice*, *19*, 36–40.

Gardezi, F., Lingard, L., Espin, S., Whyte, S., Orser, B., & Baker, G. R. (2009). Silence, power and communication in the operating room. *Journal of Advanced Nursing*, *65*, 1390–1399.

Greenberg, C. C., Regenbogen, S. E., Studdert, D. M., Lipsitz, S. R., Rogers, S. O., Zinner, M. J., & Gawande, A. A. (2007). Patterns of communication breakdowns resulting in injury to surgical patients. *Journal of the American College of Surgeons*, *204*, 533–540.

Halverson, A. L., Casey, J. T., Andersson, J., Anderson, K., Park, C., Rademaker, A. W., & Moorman, D. (2011). Communication failure in the operating room. *Surgery*, *149*, 305–310.

Interprofessional Education Collaborative Panel. (2011). *Core competencies for interprofessional collaborative practice report of an expert panel interprofessional education collaborative*. Washington, DC: Author.

Jentsch, F. G., Salas, E., Sellin-Wolters, S., & Bowers, C. A. (1995). Crew coordination behaviors as predictors of problem detection and decision making times. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 39, No. 20, pp. 1350–1353). Los Angeles, CA: SAGE.

The Joint Commission. (2015). *Sentinel event statistics data-root causes by event type (2004-2014)*. Retrieved from https://www.tsigconsulting.com/tolcam/wp-content/uploads/2015/04/TJC-Sentinel-Event-Root_Causes_by_Event_Type_2004-2014.pdf

Joint Commission Resources (2008). *Advanced lean thinking: Proven methods to reduce waste and improve quality in health care*. Oak Brook, IL: Joint Commission Resources.

Kalantari, R., Zakerian, S. A., Mahmodi Majdabadi, M., Zanjirani Farahani, A., Meshkati, M., & Garosi, E. (2016). Assessing the teamwork among surgical teams of hospitals affiliated to social security organizations in Tehran City. *Journal of Hospital*, *15*, 21–29.

Kalantari, R., Zanjirani Farahani, A., Garosi, E., Badeli, H., & Jamali, J. (2019). Translation and psychometric properties of the Persian version of oxford non-technical skills 2 system: Assessment of surgical teams' non-technical skills in orthopedic surgery wards. *The Archives of Bone and Joint Surgery*, *7*, 173–181.

Kertesz, L., Walker, C., & Maliwat-Bandigan, B. (2018). Improving communication and teamwork in the operating room. *International Journal of Nursing Care*, *2*(2), 1–10.

Keyton, J., & Beck, S. J. (2010). Perspectives: Examining communication as macrocognition in STS. *Human Factors*, *52*, 335–339.

Lingard, L., Espin, S., Whyte, S., Regehr, G., Baker, G., Reznick, R., . . . Grober, E. (2004). Communication failures in the operating room: An observational classification of recurrent types and effects. *BMJ Quality & Safety*, *13*, 330–334.

Miller, A., Weinger, M. B., Buerhaus, P., & Dietrich, M. S. (2010). Care coordination in intensive care units: Communicating across information spaces. *Human Factors*, *52*, 147–161.

Nagpal, K., Ahmed, K., Vats, A., Yakoub, D., James, D., Ashrafian, H., . . . Athanasiou, T. (2010). Is minimally invasive surgery beneficial in the management of esophageal cancer? A meta-analysis. *Surgical Endoscopy*, *24*, 1621–1629.

Neily, J., Mills, P. D., Young-Xu, Y., Carney, B. T., West, P., Berger, D. H., . . . Bagian, J. P. (2010). Association between implementation of a medical team training program and surgical mortality. *JAMA*, *304*, 1693–1700.

Parker, S. H., Wadhera, R., Wiegmann, D., & Sundt, T. (2009). The impact of protocolized communication during cardiac surgery. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 53, No. 11, pp. 684–688). Los Angeles, CA: SAGE.

Parush, A., Kramer, C., Foster-Hunt, T., Momtahan, K., Hunter, A., & Sohmer, B. (2011). Communication and team situation awareness in the OR: Implications for augmentative information display. *Journal of Biomedical Informatics*, *44*, 477–485.

Persoon, M. C., Broos, H. J., Witjes, J. A., Hendrikx, A. J., & Scherpbier, A. J. (2011). The effect of distractions in the operating room during endourological procedures. *Surgical Endoscopy*, *25*, 437–443.

Safran, D. G., Miller, W., & Beckman, H. (2006). Organizational dimensions of relationship-centered care theory, evidence, and practice. *Journal of General Internal Medicine*, *21*, 9–15.

Schiff, L., Moulder, J., Louie, M., & Toubia, T. (2016). Quality improvement of operating room communication. *Journal of Minimally Invasive Gynecology*, *23*(7), Article S58.

Seelandt, J. C., Tschan, F., Keller, S., Beldi, G., Jenni, N., Kurmann, A., . . . Semmer, N. K. (2014). Assessing distractors and teamwork during surgery: Developing an event-based method for direct observation. *BMJ Quality & Safety*, *23*, 918–929.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of information: Corrected version. *The Bell System Technical Journal*, *27*, 4–7.

Wang, J., Chellali, A., & Cao, C. G. (2016). Haptic communication in collaborative virtual environments. *Human Factors*, *58*, 496–508.

Webster, J. L., & Cao, C. G. (2006). Lowering communication barriers in operating room technology. *Human Factors*, *48*, 747–758.

Wheelock, A., Suliman, A., Wharton, R., Babu, E., Hull, L., Vincent, C., . . . Arora, S. (2015). The impact of operating room distractions on stress, workload, and teamwork. *Annals of Surgery*, *261*, 1079–1084.

Wilson, J. L., Whyte, R. I., Gangadharan, S. P., & Kent, M. S. (2017). Teamwork and communication skills in cardiothoracic surgery. *The Annals of Thoracic Surgery*, *103*, 1049–1054.

Ehsan Garosi is a PhD student of ergonomics in the Department of Occupational Health Engineering at Tehran University of Medical Sciences, Iran. He earned his MS in ergonomics from Tehran University of Medical Sciences in 2016.

Reza Kalantari is a PhD student in ergonomics at Shiraz University of Medical Sciences, Iran and a researcher at University of Jaume I, Castellon de la Plana, Spain. He received his MS in ergonomics from Tehran University of Medical Sciences in 2016.

Ahmad Zanjirani Farahani currently works in the Farahan Health Care Network, Markazi Province, Iran. He received his MS in occupational health engineering from Tehran University of Medical Sciences in 2016.

Mojgan Zuaktafi an currently works in the Department of Ergonomics, Shiraz University of Medical Sciences, Iran. She earned her MS degree in ergonomics from Tehran University of Medical Sciences in 2016.

Esmaeil Hosseinzadeh Roknabadi is a PhD student in healthcare management at Iran University of Medical Sciences, Iran. He received his MS degree in healthcare management from Tehran University of Medical Sciences in 2016.

Ehsan Bakhshi currently works at Islamabad-e Gharb Health Care Network affiliated with Kermanshah University of Medical Sciences, Kermanshah, Iran. He received his MS in ergonomics from Tehran University of Medical Sciences in 2017.

Date received: December 2, 2017
Date accepted: May 28, 2019

An Experimental Validation of Masking in IEC 60601-1-8:2006-Compliant Alarm Sounds

Matthew L. Bolton, **Xi Zheng**, **Meng Li**, University at Buffalo, The State University of New York, USA, **Judy Reed Edworthy**, University of Plymouth, UK, and **Andrew D. Boyd**, The University of Illinois at Chicago, USA

Objective: This research investigated whether the psychoacoustics of simultaneous masking, which are integral to a model-checking-based method, previously developed for detecting perceivability problems in alarm configurations, could predict when IEC 60601-1-8-compliant medical alarm sounds are audible.

Background: The tonal nature of sounds prescribed by IEC 60601-1-8 makes them potentially susceptible to simultaneous masking: where concurrent sounds render one or more inaudible due to human sensory limitations. No work has experimentally assessed whether the psychoacoustics of simultaneous masking accurately predict IEC 60601-1-8 alarm perceivability.

Method: In two signal detection experiments, 28 nursing students judged whether alarm sounds were present in collections of concurrently sounding standard-compliant tones. The first experiment used alarm sounds with single-frequency (primary harmonic) tones. The second experiment's sounds included the additional, standard-required frequencies (often called subharmonics). *T* tests compared miss, false alarm, sensitivity, and bias measures between masking and nonmasking conditions and between the two experiments.

Results: Miss rates were significantly higher and sensitivity was significantly lower for the masking condition than for the nonmasking one. There were no significant differences between the measures of the two experiments.

Conclusion: These results validate the predictions of the psychoacoustics of simultaneous masking for medical alarms and the masking detection capabilities of our method that relies on them. The results also show that masking of an alarm's primary harmonic is sufficient to make an alarm sound indistinguishable.

Application: Findings have profound implications for medical alarm design, the international standard, and masking detection methods.

Keywords: medical devices and technologies, audition, patient safety, psychophysical methods, signal detection theory

Address correspondence to Matthew L. Bolton, Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, 342 Bell Hall, Buffalo, NY 14260, USA; e-mail: mbolton@buffalo.edu.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 954–972
DOI: 10.1177/0018720819862911
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

In modern medical environments, a single patient produces hundreds of alarms per day and thus tens of thousands of alarms are generated a day in any given hospital (The Joint Commission, 2013a). Health professionals do not always respond to these alarms, and this is a very dangerous problem. The Pennsylvania Patient Safety Authority (2009) reported 194 problems (12 that resulted in fatalities) with medical personnel failing to react to telemetry monitoring alarms from June 2004 through December 2008. Similarly, 98 alarm nonresponses (five extended patient hospital stays, 13 produced "permanent loss of function," and eight ended in patient death) were documented in a Sentinel Event Alert (The Joint Commission, 2013a) that covered a period from January 2009 to June 2012. Because of these types of problems, the Emergency Care Research Institute (ECRI) has consistently named medical alarms one of the most important technological hazards to patient safety for more than a decade (ECRI Institute, 2018; Stead & Lin, 2009).

There are many reasons why humans may fail to respond to medical alarms including the number of false alarms, the lack of consistent design philosophies between alarms and medical devices, and designs that do not facilitate alarm learnability and discernibility (see reviews by Edworthy, 2013; Edworthy et al., 2018). The perceivability of the alarms in the presence of other alarms is at least partially responsible for this problem (ECRI Institute, 2014; The Joint Commission, 2013a, 2013b; Vockley, 2014).

One issue that can affect the perceivability of medical alarms is simultaneous masking. In simultaneous masking, limits of the human sensory system prevent humans from hearing one or more concurrent sounds (Fastl & Zwicker,

2006). A number of researchers have generally speculated that simultaneous masking could be a problem with medical alarms (Edworthy & Hellier, 2005, 2006; Edworthy & Meredith, 1994; Konkani, Oakley, & Bauld, 2012; Meredith & Edworthy, 1995; Patterson, 1982; Patterson & Mayfield, 1990). This is because medical alarms are often represented as melodies of tonal sounds, including alarms that are compatible with the International Electrotechnical Commission's (IEC) international standard (IEC 60601-1-8:2006/AMD1:2012, 2012). This makes them especially prone to simultaneous masking (Bosi & Goldberg, 2003; Fastl & Zwicker, 2006). There is also empirical evidence that simultaneous masking does occur for medical alarms in modern hospitals. Momtahan, Hetu, and Tansley (1993) analyzed 49 medical alarms and found 25 pairs in which one could be completely masked by the other. Toor, Ryan, and Richard (2008) discovered several instances where high priority alarms could be masked by lower priority operating room sounds including other alarms, telephone rings, and beeper sounds. Both of these studies involved recording sounds in a medical environment and then using the psychoacoustics of simultaneous masking (mathematical formulations that predict whether simultaneous masking occurs based on the volumes and frequencies of the sounds; Bosi & Goldberg, 2003) to identify pairs of alarms where masking could occur.

Despite these findings, medical alarm safety has mostly focused on other problems (Edworthy, 2013). This is likely due to the complexity of simultaneous masking. Masking can manifest as a result of multiple simultaneously sounding alarms (not just pairs) and may only occur for particular timings of the overlaps between the alarms. It is thus almost impossible for analysts to experimentally determine how masking could manifest in alarm configurations. Given the sheer number of medical alarms and possible different overlaps between them in a given hospital (The Joint Commission, 2013a), it is likely that masking is an important factor in alarm nonresponse.

To address this situation, we developed a computational method (Bolton, Edworthy, & Boyd, 2018; Bolton, Edworthy, Boyd, Wei, & Zheng, 2018; Bolton, Hasanain, Boyd, & Edworthy, 2016; Hasanain, Boyd, & Bolton, 2016, 2014; Hasanain, Boyd, Edworthy, & Bolton, 2017) that uses the psychoacoustics of simultaneous masking and model checking. Model checking is a formal method that allows an analyst to automatically, mathematically prove properties against models of concurrent systems (a process called formal verification; Clarke, Grumberg, & Peled, 1999). In our method, an analyst models the behavior of alarms and runs model checking to prove if the represented alarms can ever mask each other. This method has been used to analyze real medical alarm configurations (Bolton et al., 2018; Bolton et al., 2016; Hasanain et al., 2017) and the reserved alarm sounds of the IEC 60601-1-8 international standard (Bolton et al., 2018).

This method is powerful and offers unprecedented masking detection capabilities. However, the method has limitations. First, like the experimental results presented by Momtahan et al. (1993) and Toor et al. (2008), the method relies on the psychoacoustics of simultaneous masking. Although these psychoacoustics have been well tested over the years (Bosi & Goldberg, 2003), they have not been explicitly experimentally validated for medical alarm sounds. Second, many tonal medical alarms are consistent with the IEC 60601-1-8 standard. This means that they contain a primary harmonic (frequency) as well as several additional harmonics (usually the minimum of 4) that are multiples of the primary that are at lower volumes. Although our method is capable of accounting for the masking effects of both primary and additional harmonics, including the additional harmonics can require orders of magnitude more computational time. Thus, if the masking of the primary harmonics was critical to alarm perceivability irrespective of the additional harmonics, this would profoundly improve the usefulness and relevance of our method.

We addressed both of these issues by conducting two signal detection theory (SDT) experiments. In the first, we validated the ability of our method to predict masking between primary harmonics of IEC 60601-1-8 medical alarm sounds. In the second, we assessed how well predictions about masking between the primary harmonics affect the perceivability of

TABLE 1: IEC 60601-1-8 Alarm Tone Characteristics

Tone Characteristic	Value Range
Primary frequency (Hz)	[150, 1000]
Primary frequency volume (dBs)	v
Maximum primary tone volume difference (dBs)	10
Minimum number of additional harmonics	4
Additional harmonics frequency (Hz)	[300, 4,000]
Additional harmonic volume (dBs)	[$v - 15$, $v + 15$]
Duration (s)	[0.075, 0.25]

Note. dBs = decibels.

alarms with a full set of IEC 60601-1-8-required additional frequencies.

REVIEW OF RELEVANT LITERATURE

Below we provide background on the alarms of IEC 60601-1-8, the psychoacoustics of simultaneous masking that are used by our method to predict masking, and the SDT experimental paradigm that we use in our research.

IEC 60601-1-8

The IEC 60601-1-8 international medical alarm standard is widely used across the medical industry. It was created to improve alarm discernibility and identification. As part of this, it provides instructions for designing new alarm sounds, which typically manifest as melodies (sequences) of tones separated by pauses. There are many details in the standard. For the work presented in this paper, we are primarily concerned with the specific requirements of the individual tones that compose alarm melodies. Each tone in a melody has a single primary frequency. It also has several additional harmonics (additional frequencies) designed to make the alarms more tonally rich and help listeners localize alarm sources. The standard does not require specific frequencies, volumes, and timings of the tones in alarm melodies. Rather, it provides ranges of acceptable values. These are summarized in Table 1.

A number of issues have been identified with the melodic alarm sounds prescribed in the standard that compromise the standard's goal of making alarms discernable and identifiable (Edworthy et al., 2018). In this work, we are particularly concerned with the effect simultaneous masking has on alarm audibility.

Masking and the Psychoacoustics of Simultaneous Masking

Auditory masking describes a number of different phenomena where a sound is rendered inaudible due to the presence of one or more other (masking) sounds. For example, pressure waves of sounds can physically interact to cancel each other out or a given sound can be indistinguishable from environmental noise. In this work, we focus on simultaneous masking. This occurs when similar, simultaneous sounds render one or more imperceptible due to the way that the sounds affect the sensitivity of the human sensory system.

Our method uses the psychoacoustics of simultaneous masking to make predictions about whether any given alarm in a configuration will be audible. The psychoacoustics of simultaneous masking mathematically describe how the volumes and frequencies of sounds produce masking. In particular, the psychoacoustics are based on how masking sounds (*maskers*) stimulate the sensors of the basilar membrane: the spiral-shaped physical structure in the human inner ear that is responsible for the ability of humans to distinguish between sounds (Ambikairajah, Davis, & Wong, 1997; Baumgarte, Ferekidis, & Fuchs, 1995; Bosi & Goldberg, 2003; Brandenburg & Bosi, 1997; Brandenburg & Stoll, 1994; Schroeder, Atal, & Hall, 1979). This raises the absolute threshold (in decibels [dBs]) that the volume of another sound (a potential *maskee*) must exceed to be perceivable (Bosi & Goldberg, 2003).

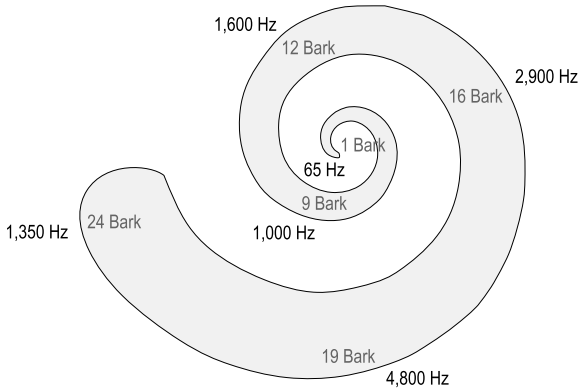


Figure 1. Depiction of how peak stimulation of sounds in Hertz occurs at different Bark locations along the basilar membrane.

These psychoacoustics render frequencies on the Bark scale (Zwicker & Feldtkeller, 1967): a scale that maps a frequency in Hertz to a position on the basilar membrane where that frequency most powerfully stimulates the receptors (see Figure 1). A frequency in Hertz (f_{sound}) is converted into Barks using Equation 1.

$$z_{\text{sound}} = 13 \cdot \arctan\left(76 \cdot \frac{f_{\text{sound}}}{100000}\right) + 3.5 \cdot \arctan\left(\left(\frac{f_{\text{sound}}}{7500}\right)^2\right) \quad (1)$$

The "masking curve" calculates how a given masker shifts the absolute threshold of hearing with Equation 2.

$$\text{curve}_{\text{masker}}\left(z_{\text{maskee}}\right) = \text{spread}_{\text{masker}}\left(\delta z\right) v_{\text{masker}} - \Delta \quad (2)$$

In this, v_{masker} is the masker's volume in decibels and δz is calculated using

$$\delta z = z_{\text{maskee}} - z_{\text{masker}} \quad (3)$$

where z_{maskee} and z_{masker} are the Bark scale frequencies of the maskee and masker, respectively. Furthermore, the $\text{spread}_{\text{masker}}$ (Equation 2) function models how the magnitude/volume of the masking threshold changes with respect to δz. Finally, Δ is the minimum difference between the volumes of the masker and maskee that can result in masking.

There are multiple formulations of the psychoacoustic spreading function and Δ based on the characteristics of the masking and masker sounds. In this research, we use the formulation

$$\text{spread}_{\text{masker}}\left(\delta z\right) = \begin{cases} -17 \cdot \delta z + 0.15 \cdot v_{\text{masker}} \cdot \left(\delta z - 1\right) \cdot \theta\left(\delta z - 1\right) & \text{for } \delta z \geq 0 \\ -\left(6 + 0.4 \cdot v_{\text{masker}}\right) \cdot |\delta z| - \left(11 + 0.4 \cdot v_{\text{masker}} \cdot \left(|\delta z| - 1\right)\right) \cdot \theta\left(|\delta z| - 1\right) & \text{otherwise} \end{cases} \quad (4)$$

where $\theta(x) = 1$ for $x \geq 0$ and $\theta(x) = 0$ otherwise,

$$\Delta = 6.025 + 0.275 \cdot z_{\text{masker}} \text{ dB} \quad (5)$$

These were used because they are universally regarded as the most accurate for modeling tonal sounds (Ambikairajah et al., 1997; Bosi & Goldberg, 2003; Brandenburg & Stoll, 1994). Figure 2 illustrates the shape of the masking curve described by Equation 2.

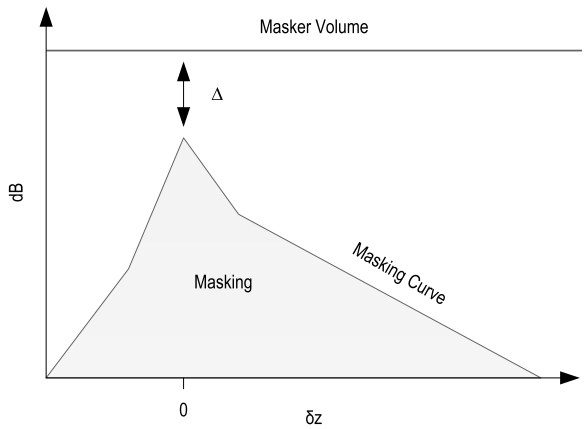


Figure 2. The masking curve shape dictated by Equations 2 to 5. dB = decibel.

Furthermore, the combined masking threshold of multiple concurrent sounds can be greater than a simple sum of the effect of individual maskers (Bosi & Goldberg, 2003; Humes & Jesteadt, 1989). This effect is called additive masking. Because masking levels are measured in decibels (a logarithmic scale), these are transformed to the power scale to allow for arithmetic operations. A volume in decibels (v) can be converted to the power scale using Equation 6.

$$\text{power}(v) = 10^{\frac{v}{10}} \tag{6}$$

Then, for a given potential maskee and N potential maskers, the absolute value of hearing adjusted for the additive effect of masking (in decibels) is calculated using Equation 7.

$$\text{power}\left(\text{mthresh}_{\text{maskee}}\right) = \text{power}\left(\text{abs}_{\text{maskee}}\right) + \left(\sum_{n=1}^{N} \text{power}\left(\text{curve}_{\text{masker}_n}\left(z_{\text{maskee}}\right)\right)^{\alpha}\right)^{\frac{1}{\alpha}} \tag{7}$$

In this, α is a positive constant (Green, 1967); $\text{abs}_{\text{maskee}}$ is the unaltered absolute threshold of hearing (in decibels) at the maskee's frequency. This, using the maskee's frequency in Hertz, is described using Equation 8 (Terhardt, 1979).

$$\text{abs}_{maskee} = 3.64 \cdot \left(\frac{f_{\text{maskee}}}{1000}\right)^{-0.8} - 6.5 \cdot \text{e}^{-0.6\left(\frac{f_{\text{maskee}}}{1000-3.3}\right)^2} + 10^{-3} \cdot \left(\frac{f_{\text{maskee}}}{1000}\right)^4 \tag{8}$$

These psychoacoustics have been used successfully to predict masking for normal human hearing for decades (Bosi & Goldberg, 2003). They were employed by researchers to identify when masking could occur for sounds recorded in medical environments (Momtahan et al., 1993; Toor et al., 2008). They were also the basis for lossy audio compression techniques (like those used in MPEG [Moving Picture Experts Group] formats; Bosi & Goldberg, 2003), digital audio compression methods that allow reductions in the size of audio files by removing audio data that is predicted to be masked.

SDT

SDT models the detection of an event in a noisy environment. In a human judgment context, this captures both the state of the world (whether there is signal in the presence of noise or just noise) and the human's response ("Yes" there is a signal or "No" there is no signal). Based on this representation, there are four possible classifications of the outcome (Figure 3). Two of these are correct. If the judge says "Yes"

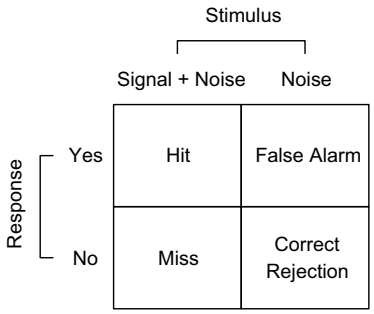


Figure 3. A matrix describing how the different outcomes can manifest based on a "Yes" or "No" human response to a stimulus that is either signal or noise.

when there is signal, the outcome is a hit. If the judge says "No" when there is only noise, the outcome is a correct rejection. Two of the outcomes are incorrect. If the judge says "Yes" when there is only noise, the outcome is a false alarm. If the judge says "No" when there is a signal, the outcome is a miss.

When a human performs a signal detection task and makes multiple judgments in response to different states of the world, rates can be calculated for each of the outcomes:

Hit Rate:

$$H = \frac{\text{No. of hits}}{\text{No. of signal events}}, \tag{9}$$

Miss Rate:

$$M = \frac{\text{No. of misses}}{\text{No. of signal events}} = 1 - H, \tag{10}$$

False Alarm Rate:

$$F = \frac{\text{No. of false alarms}}{\text{No. of noise events}}, \tag{11}$$

Correct Rejection Rate:

$$C = \frac{\text{No. of correct rejections}}{\text{No. of noise events}} = 1 - F. \tag{12}$$

Note that because of the inverse relationships between hits and misses and between false alarms and correct rejections, analysts will typically only discuss results from one rate from each pair. For example, in the presented work, we only talk about miss and false alarm rates.

Two additional measures for modeling human judgment are typically calculated from the above rates: sensitivity and response bias (or simply bias). Sensitivity captures the judge's ability to distinguish signal from noise. Response bias is a measure of whether a judge is more likely to respond one way or another.

When the signal and noise can be assumed to be normally distributed with equal variance, sensitivity is the distance between the means of the signal and the noise distributions. The response bias is the likelihood ratio that a response of "Yes" is due to the presence of signal as opposed to noise alone. However, for many judgment tasks, the distributions of signal and noise may not be normally distributed (as will be the case in the experiments presented in this paper) or the distributions may be unknown. Thus, there are nonparametric measures for computing sensitivity and response bias (Macmillan & Creelman, 1990; See, Warm, Dember, & Howe, 1997). In this work, we use the nonparametric calculations that have been shown to be appropriate in human subject experiments (See et al., 1997).

A', based on concepts introduced by Pollack and Norman (1964), calculates nonparametric sensitivity by approximating the area under a receiver operating characteristic (ROC) curve defined by the observed hit (H; Equation 9) and false alarm (F; Equation 11) rates (Snodgrass & Corwin, 1988):

$$A' = \begin{cases} 0.5 + \dfrac{(H-F)\cdot(1+H-F)}{4\cdot H\cdot(1-F)} & \text{if } H \geq F \\ 0.5 + \dfrac{(F-H)\cdot(1+F-H)}{4\cdot F\cdot(1-H)} & \text{otherwise} \end{cases}. \tag{13}$$

This produces a value between 0 and 1, where a higher value indicates that the judge was more sensitive (more readily able to distinguish between signal and noise).

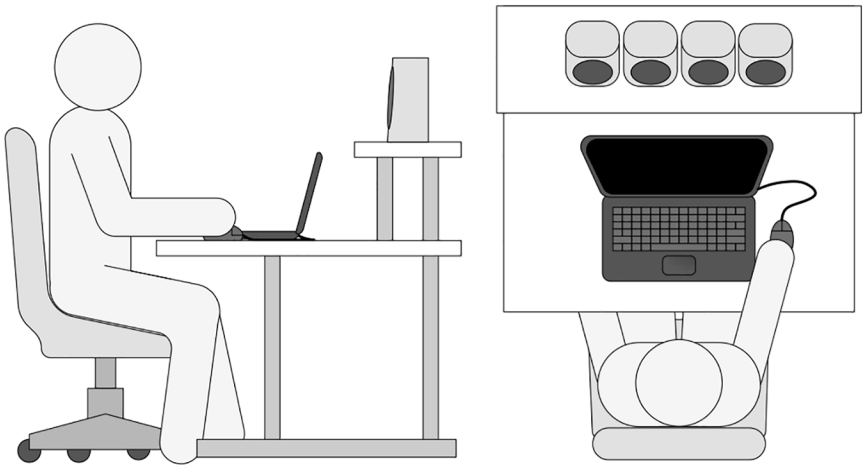


Figure 4. The physical apparatus setup used in the reported experiments. Both depict ((a) in profile and (b) from above) a participant sitting in front of a computer desk on which a laptop computer, a computer mouse, and four speakers were placed.

B''_D, which was introduced by Donaldson (1992), is a nonparametric measure of response bias that is also based on the geometry of the ROC curve:

$$B''_D = \frac{(1-H)\cdot(1-F) - H\cdot F}{(1-H)\cdot(1-F) + H\cdot F}. \qquad (14)$$

A B''_D bias will range between −1 and 1, where a negative value indicates that the judge is more likely to say no (has a conservative bias), a positive value indicates that the judge is more likely to say yes (has a liberal bias), and a value of 0 indicates that the judge is just as likely to say one or the other.

EXPERIMENT 1

In our first experiment, we used a SDT procedure to assess how well the psychoacoustics of simultaneous masking that are used in our method predict the ability of humans to perceive the primary harmonics of alarm sounds from IEC 60601-1-8.

Method

Participants. A power analysis revealed that 80% power was achieved for detecting a moderate effect size ($d = 0.55$) with a two-tailed paired t test with 28 participants. Thus, 28 participants were recruited for this study. Nursing students from the University at Buffalo were used as the participant pool because it constituted members of the actual population that will experience medical alarm sounds in a natural environment. Twenty-one of the recruited students were female and seven were male. The experiment did not control for musical experience because we could find no research showing that musical ability had any impact on masking.

Materials and apparatus. The experiment was run in the Usability Laboratory at the University at Buffalo, a controlled, quiet, evenly lit environment. It was administered on a laptop computer resting on a computer desk (see Figure 4) in front of which a participant would sit. The laptop computer was connected to an external USB, 7.1 sound card. Four single-driver computer speakers were connected to the sound card so that each speaker only output sounds sent to a single channel of the sound card. The speakers were placed in line with each other on an elevated platform behind the laptop. The laptop computer was also connected to an optical computer mouse that the participant could use to interact with the software that administered the experiment.

Figure 5. The interface to the software used to administer the experiment and collect participant responses. This was always displayed in full screen so that the user could not see or interact with the Excel spreadsheet running in the background.

The software used for administering the experiment was constructed specifically for this study. This was implemented as a Visual Basic for Applications program within a Microsoft Excel spreadsheet. This software was able to examine the experimental design (which was stored in the spreadsheet), administer a given participant's experiment according to it, collect user responses, and store them in a separate Excel sheet. The interface that the software used for administering the experiment is shown in Figure 5. This told a participant which trial they were on, out of the total number of trials. It also gave participants instructions for how to perform the trial.

In a given trial, participants were charged with determining whether a judgment sound was present in a test sound. The judgment sound represented a single alarm sound that was always played on the rightmost speaker (Figure 4). The test sound constituted a simulation of the simultaneous sounding of multiple alarm sounds (between one and three) from different devices. Thus, the alarm sounds of the test sound were each played on one of the three left-most speakers (one sound per speaker; Figure 4). When interacting with a trial in our software, a participant would first click on the "Play the Judgment Sound" button to play the judgment sound. Participants would then click on the "Play the Test Sound" button to play the test sound (Figure 4). Participants were allowed to play either of these sounds (one at a time) as many times as they wanted to until they felt like they could render a judgment. When participants were ready, they would indicate whether or not they thought the judgment sound was present in the test sound by clicking on the "Yes" (indicating they thought the sound was present) or "No" (indicating that they thought the sound was not present) radio buttons. When participants were satisfied with their answer, they would click on the "Next" button. The interface would then present a dialogue box that would ask participants if they wanted to confirm their answer. If participants pressed a "No" button, they would stay on the

TABLE 2: The Frequencies Used in Tones Found in Judgment and Test Sounds

Scientific Pitch Notation	Frequency (Hz)
C_4	261.63
$C\#_4$	277.18
D_4	293.66
$D\#_4$	311.13
E_4	329.63
F_4	349.23
$F\#_4$	369.99
G_4	392.00
$G\#_4$	415.30
A_4	440.00
$A\#_4$	466.16
B_4	493.88
C_5	523.25

current trial. If they pressed a "Yes" button, they would go to the next trial. Whenever this dialogue box was being displayed, the software played brown noise (signal noise naturally produced by Brownian motion; Vasseur & Yodzis, 2004) from the speakers to give participants a "palate cleanser" between trials.

A sound level meter, positioned at the ear position of a participant, was used to calibrate the laptop and speakers so that volumes matched the levels specified by the experiment.

A given trial was a pair of sounds: the judgment sound (a single alarm sound played on a single speaker in the apparatus) and the test sound (a collection of one to three alarm sounds, each played synchronously on a separate speaker in the apparatus). All of the sounds were designed to be consistent with the requirements of single tones from IEC 60601-1-8-compliant alarm sounds (Table 1), with only the primary harmonics. Each of the sounds was 0.25 s long. The tone in the judgment sound was 70 dB. One of the tones in all of the test sounds was 70 dB. The other tones in the test sound were 85 dB. If the judgment sound was in the test sound, the judgment sound was always the 70 dB sound in the test sound. These volumes were used because they were allowable by the standard (which specifies variations in volumes of any alarm sounds at the same priority be within 15 dB of each other; Table 1), are consistent with alarm volumes used in the field, and were not loud enough to cause hearing problems when combined together in the experiment. The set of frequencies used for tones were based on piano notes that fit within the allowable range of the standard (and are used to formulate the reserved alarm sounds in the standard; see Table 2). The frequencies used for the tones of the test sound were always different from each other.

Independent variables. There were two independent within-subject variables in the experiment that enabled the use of a SDT experimental design. First, a trial was either a signal trial or a noise trial. In a signal trial, the judgment sound was one of the sounds output in the test sound. In a noise trial, the judgment sound was not part of the test sound. Second, a trial could either contain masking (where the 70 dB tone was masked by the other tones in the test sound according to the psychoacoustics of simultaneous masking) or not (where, according to the psychoacoustics, none of the tones in the test sound would be masked). In trials that were both signal and masking, the test sound would contain the judgment sound and the judgment sound would be the one predicted to be masked.

Dependent measures. For each trial, participants would indicate whether they thought the judgment sound was present in the test sound. This "Yes" (the judgment sound was present) or "No" (the judgment sound was absent) response was the only dependent measure in the experiment.

Procedure. In the experiment, a participant was admitted to the lab and sat in front of the apparatus as shown in Figure 4. The participant was then given an informed consent document which they read and signed. After this, participants were read instructions that told them how to interact with the software interface (Figure 5) to administer the experiment. The participants were given a copy of the instructions for their reference. The participant then interacted with the software's interface to administer training and the experiment. When the experiment was completed, participants were given a US$20 Amazon gift card.

Training. Before the proper start of the experiment, all participants experienced the same 18 training trials that were designed to introduce them to the judgment task. This was done by presenting trials in blocks. All trials and blocks were always presented to participants in the same order. The first block of four trials were signal trials that did not contain masking. The second block of four trials were noise trials that also did not contain masking. The third block of four trials were signal trials that did contain masking. For all three of these blocks, dialogue boxes introduced the blocks and told patients whether he or she should or should not hear the judgment sound in the test sound. In the final block of six trials, the trials were a random ordering of two signal trials with masking, two signal trials without masking, one noise trial without masking, and one signal trial without masking. In this final block, participants were told it was up to them to determine if the test sound was in the judgment sound. Across all of the training trials, participants were given feedback, via a dialogue box, about the accuracy of each judgment after it was made.

Experimental design. Following training, each participant experienced the same 200 experimental trials. These trials were grouped in a single block and were arranged consistently with the standards for nonparametric, human subjects, SDT designs as outlined by McNicol (2005), who recommended 50 masking and 50 noise trials for each experimental condition considered in an experiment. As per these standards, trials contained 100 masking trials and 100 trials without masking, where there were 50 signal and 50 noise trials in each 100-trial designation. All 200 trials were presented to each participant in a unique, randomly generated order. In signal trials, the speaker on which the judgment sound was played as part of the test sound was counterbalanced between trials.

The number of tones included in the trial's test sound could vary. In masking trials, test sounds could have either two or three tones (there were equal numbers of masking trials with each number of tones). In trials without masking, test sounds could have between one and three tones (there were equal numbers of nonmasking trials with each number of tones). Test sounds in masking trials were not allowed to have one tone because simultaneous masking could not occur in such a situation. Test sounds were allowed to have one tone in trials without masking because it could provide a nonmasking condition that was perceptually comparable to a masking condition with two tones.

Data analysis. Because simultaneous masking theoretically makes masked alarms inaudible, we hypothesized that we would observe a significantly higher miss rate (M) for the masked trials than the unmasked ones. Due to the nature of SDT rates (Equations 9 to 12), this would correspond to a significantly lower hit rate (H) for masked trials than the unmasked trials. Because the presence or absence of masking should not affect a human's tendency to say "Yes" in noise trials, we did not hypothesize a significant difference in false alarm rates (F) (and thus correct rejection rates; C) between masked and unmasked trials.

Because the inability to hear alarms would suggest a drop in human sensitivity, we hypothesized that humans would exhibit a lower sensitivity that was significant for masked trials than for unmasked ones. We did not hypothesize that the presence of masking would affect participant response bias.

To test these hypotheses, we analyzed each participant's responses in accordance with the SDT measures discussed in the background section. First, for each participant, the masking and nonmasking trials were analyzed separately and used to compute each of the SDT rates (H, F, M, and C; Equations 9 to 12) and their associated nonparametric measures of sensitivity (A'; Equation 13) and bias (B''_D; Equation 14). Then, we used paired t tests to compare M, F, A', and B''_D across participants. For M and A', because we hypothesized a direction to differences, one-tailed tests were used. For the other measure, because no direction of difference was hypothesized, two-tailed tests were used. Ultimately, statistical significance was assessed at an alpha level of .05 that was Bonferroni adjusted for the 10 t tests performed for the research presented in this paper. This ultimately resulted in an adjusted significance level of $0.05/10 = 0.005$. Effect sizes of these tests were computed using a Cohen's d.

Note that to ensure that the assumptions for the *t* tests were valid, in all cases, an Anderson–Darling test was conducted to assess the normality of the difference between the paired rates of participants.

Results

The results of the comparisons of miss and false alarm rates (*M* and *F*, respectively) are reported in Figure 6. These analyses showed that miss rate (*M*) was significantly higher for masking trials than for trials without masking. There was no significant difference between false alarm rates (*F*) between masking and nonmasking trials.

The sensitivity (A') and bias (B''_D) results and statistics comparing them are reported in Figure 6. These analyses showed that sensitivity was significantly lower for masking trials than nonmasking trials. This means that people had a more difficult time distinguishing between signal and noise when masking was predicted than when it was not. On average, bias measures were positive. This indicates that participants tended to say "Yes" more often than they said "No." People tended to say "Yes" more often for masking trials than for nonmasking trials. This difference would have met a .05 significance level ($p = .008$); however, this failed to meet the adjusted level of statistical significance.

Discussion

These results are consistent with our hypotheses. We found that participants made more misses when the test sounds were masked than when they were not. In the nonmasking condition, participants had misses only roughly one third of the time while, in the masking condition, participants made misses on average 48.1% of the time, which is extremely close to 50% (which would be expected by random guessing). Conversely, there was no significant difference in false alarm rates between the two conditions, which happened roughly 30% of the time. Furthermore, participants had a lower sensitivity for masking trials than for nonmasking ones. Collectively, these results suggest that participants clearly had more trouble distinguishing between signal and noise in the masking condition than the nonmasking one, and that this was predominantly due to the fact that masking makes it more likely that humans will miss alarms. This is an important result because it validates that the psychoacoustics of simultaneous masking used in our method are able to accurately predict whether or not masking will contribute to alarm perceivability.

It is slightly concerning that the judgment error rates observed outside of the masking miss condition occurred roughly one third of the time and were not closer to 0. This is likely due to the fact that the judgment task was difficult and that there are higher perceptual, attentional, and cognitive factors that will influence it. Implications of this are explored in greater depth in the general discussion.

The results on bias did not strictly violate our hypothesis that there would be no statistically significant difference between the masking and nonmasking conditions. There does appear to be a trend that people were biased toward saying "Yes" more often in masking trials than in nonmasking ones. It is not entirely clear why this occurred. This will be explored in greater depth in the general discussion.

EXPERIMENT 2

Experiment 1 provided evidence for the validity of the psychoacoustics of simultaneous masking. However, because this experiment did not include any additional harmonics, it is not clear whether these results would generalize to complete alarm sounds as specified in the international standard (see Table 1). Thus, the second experiment we conducted was designed to see how well the masking of alarm sound primary harmonics affects the perceivability of more complex alarm sounds that include the requisite additional harmonics dictated by the standard (see Table 1).

Method

Experiment 2 was an almost identical replication of Experiment 1. It was performed with 28 new nursing student participants (this time with 24 females and four males) with the same apparatus, methods, and experimental design. There were two important differences.

First, while the alarm sounds represented the same set of 200 trials from the first experiment, the versions of the sounds used in Experiment 2

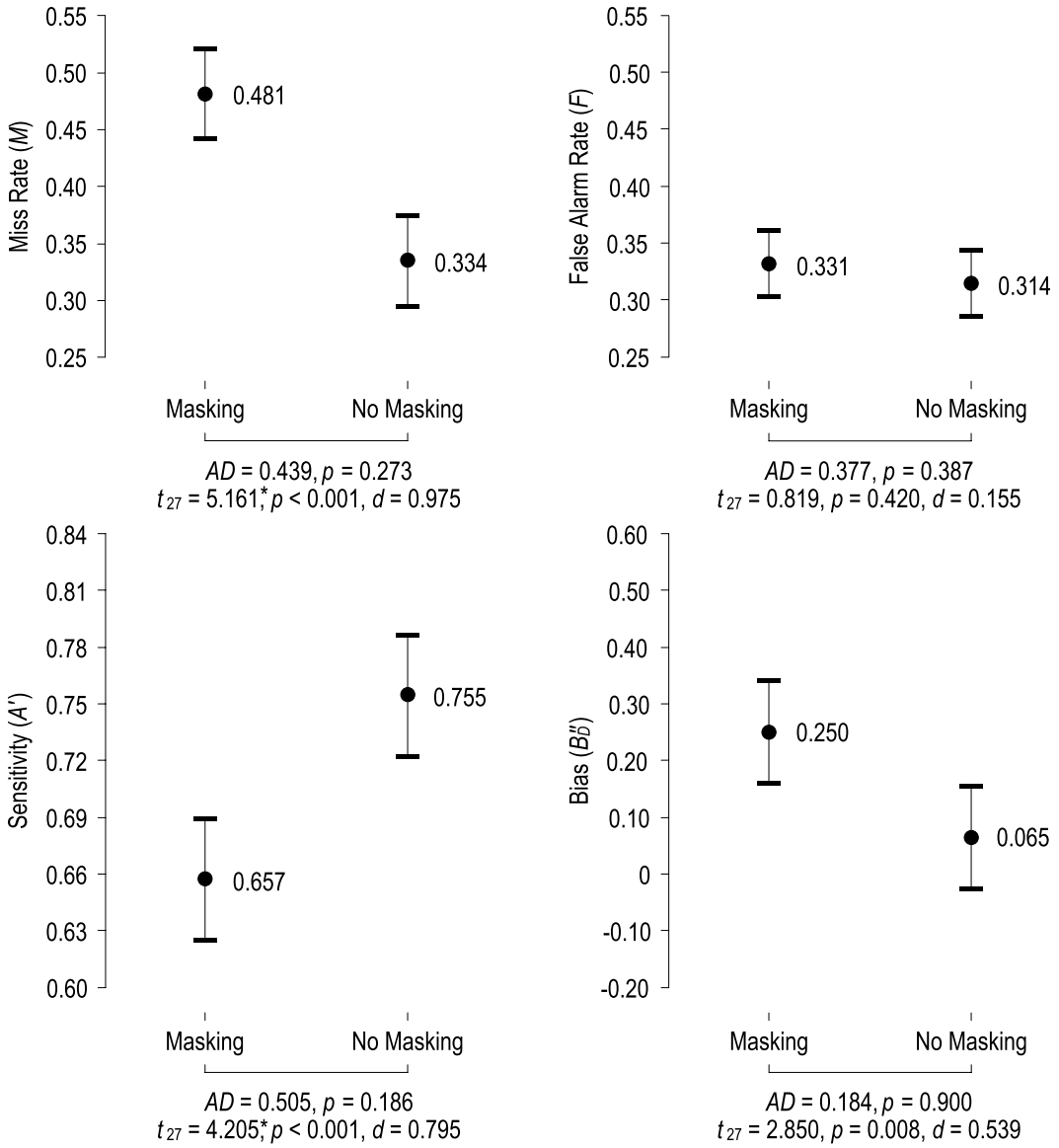


Figure 6. Means (labeled circles) and 95% within-subject confidence intervals (horizontal bars; Cousineau & O'Brien, 2014) for miss rates (*M*), false alarm rates (*F*), sensitivity (*A'*), and bias (B''_D) for both the masking and nonmasking conditions observed during Experiment 1. Rates are presented with Anderson–Darling statistics that indicate that differences between the paired rates of participants followed a normal distribution. Rates are also presented with paired *t* test results and their corresponding Cohen's *d* effect size. Statistical significance is indicated with "*."

were extended to include four additional harmonics that played concurrently with the original primary harmonic of the sound. In all cases, these additional harmonics were computed as being 3, 5, 7, and 9 times the frequency of the primary harmonic (whole number multiples are typically used to avoid dissonance in the complex sound). Each additional harmonic had a volume 15 dB lower than the primary one. These parameters made the alarm sounds compliant

with the standard (see Table 1) and were consistent with common recommendations for accomplishing this (Thompson, 2010).

The second difference from Experiment 1 came in the data analysis. While the results of Experiment 2 were evaluated using the same methods as Experiment 1, we also used standard (nonpaired) two-tailed *t* tests to determine if there were significant differences between comparable measures (*M, F, A′*, and B''_D) between the experiments. This allowed us to assess whether the inclusion of the additional harmonics improve or reduce alarm perceivability in both the presence and absence of masking.

Results

Results and statistics for the miss rate (*M*) and false alarm rate (*F*) analyses are shown in Figure 7. These showed that miss rate was significantly higher for the masking condition than for the nonmasking one and that there were no statistically significant differences in false alarm rates.

The results of the sensitivity (*A′*) and bias (B''_D) analyses are also shown in Figure 7. These showed that there were significant differences between sensitivity and bias. On average, participants were significantly less sensitive in the masking condition than in the nonmasking condition. Conversely, participants had a significantly higher bias (and thus tended to say "Yes") more often in the masking condition.

The comparison of these SDT statistics to the comparable ones from Experiment 1 (see Figure 8) revealed that there were no significant differences between any of them.

Discussion

The results for Experiment 2 effectively replicated the results seen for Experiment 1. It produces comparable values between the computed SDT measures and none of the comparable measures' difference was statistically significant. These results show that the inclusion of additional harmonics does not affect the overall perceivability of alarms for any of the experimental conditions. Given the comparable rates and sensitivities across the masking and nonmasking conditions, this means that the additional harmonics neither counteract the effect of masking nor do they help improve the overall perceivability of the alarms. This is a compelling result that will be discussed further in the next section. It is important to note that because the frequencies of the additional harmonics were obtained by multiplying the primary harmonic by whole numbers, it is extremely unlikely that any of these harmonics would be masked due to the bark distances this multiplication creates. Thus, this effect is not due to simultaneous masking. It is our hypothesis that the masking of the primary harmonic reduces the salience of the alarms such that the additional frequencies are not enough for people to identify them. This will need to be investigated more deeply in future research.

The only slight discrepancy in the results between the two experiments was seen in the response bias measures, which did exhibit a significant difference in Experiment 2 (only a nonsignificant trend was seen in Experiment 1). As with Experiment 1, it is not entirely clear why participants would tend to say "Yes" in the masking condition. This is discussed more in the next section.

GENERAL DISCUSSION AND CONCLUSION

This research used human subject experiments to validate that the psychoacoustics of simultaneous masking are able to predict the perceivability of medical alarm sounds. To the best of our knowledge, this is the first research to empirically show that masking is a problem for the current IEC 60601-1-8 alarms. Furthermore, our research showed that the masking effect is strong enough to reduce the audibility of IEC 60601-1-8-compliant alarms by a statistically significant amount, even with the inclusion of the requisite additional harmonics. These are powerful results because they mean that the psychoacoustics of simultaneous masking can be used to make predictions about whether people will be able to hear alarms from the IEC 60601-1-8 international standard and that this can be done with only the primary harmonics of the alarms.

Our results are of import to our method (Bolton et al., 2018; Bolton et al., 2018; Bolton et al., 2016; Hasanain et al., 2014; Hasanain

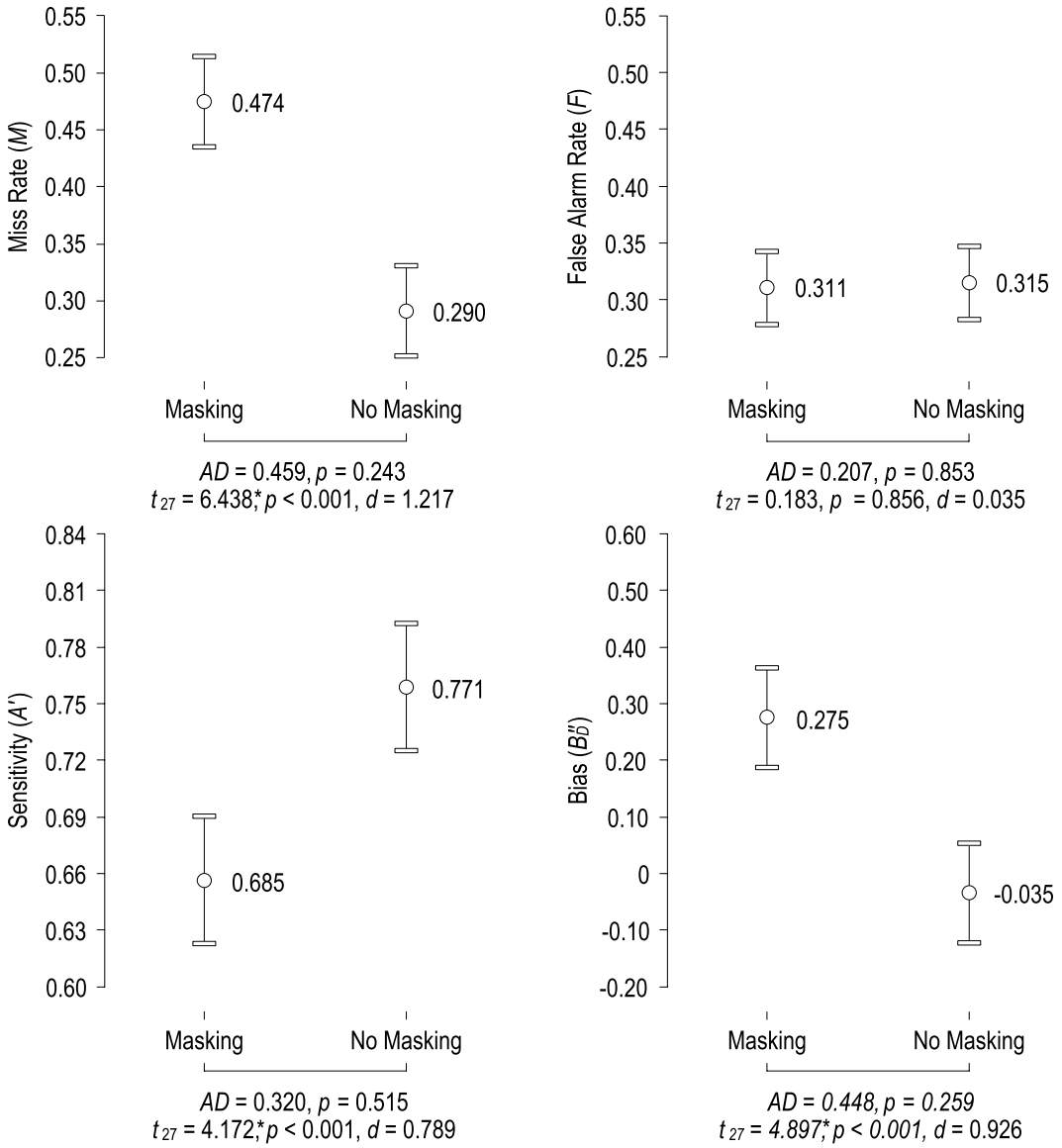


Figure 7. Means (labeled circles) and 95% within-subjects confidence intervals (horizontal bars) for miss rates (*M*), false alarm rates (*F*), sensitivity (*A*′), and bias (B''_D) for both the masking and nonmasking conditions observed during Experiment 2. Rates are presented with Anderson–Darling statistics that indicate that differences between the paired rates of participants followed a normal distribution. Rates are also presented with paired *t* test results and their corresponding Cohen's *d* effect size. Statistical significance is indicated with "*."

et al., 2016; Hasanain et al., 2017), which, in turn, has important implications for alarm design and masking in health care environments. First, by validating the predictive capabilities of the psychoacoustics that our method uses, we enable the predictive power of our method to be used effectively to design and evaluate medical alarms and its use in our ongoing effort to evaluate and improve the international medical alarm standard (Bolton et al., 2018). This has the potential to improve the perceivability of medical alarms across the industry and thus improve

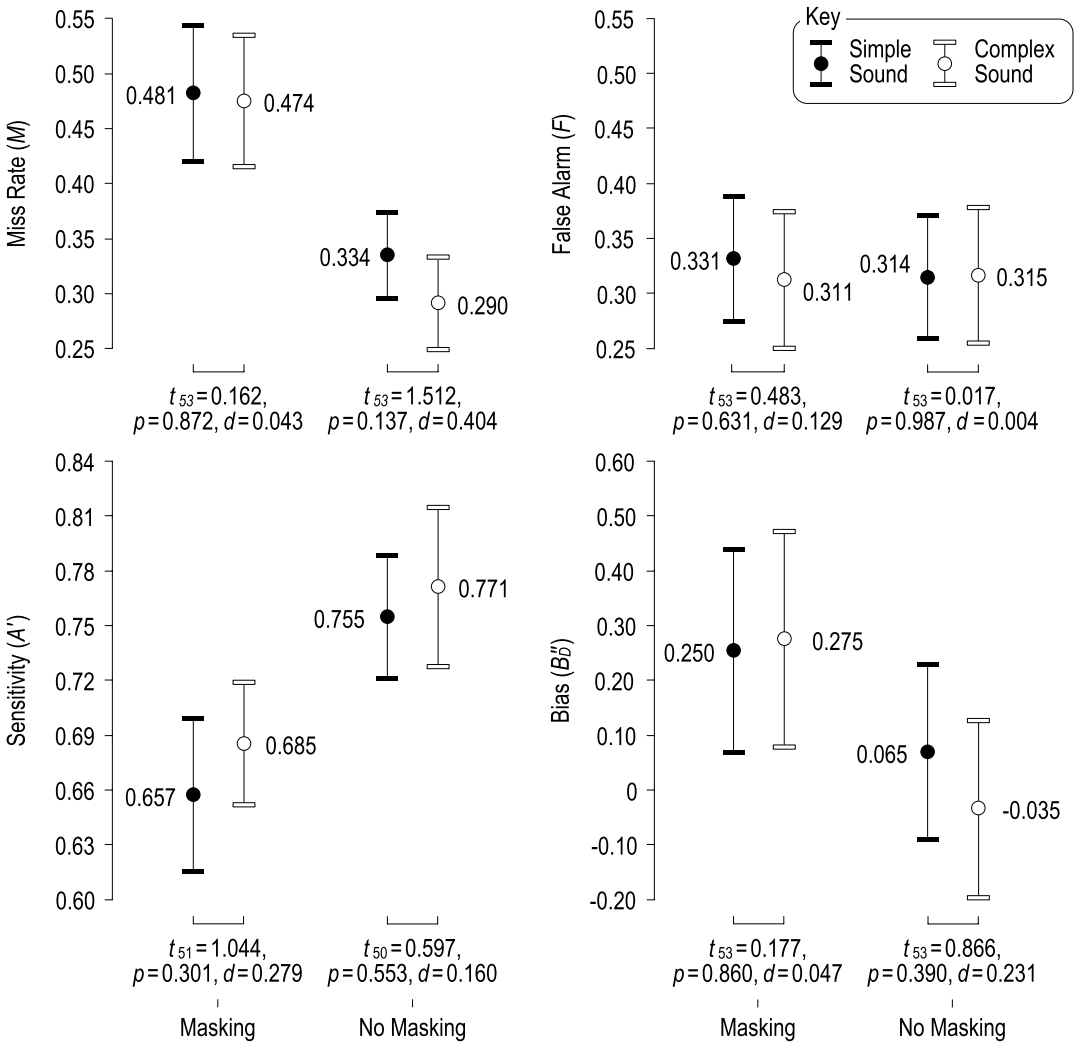


Figure 8. Comparisons of miss rate (*M*), false alarm rate (*F*), sensitivity (*A'*), and bias (B''_D) values measured in Experiments 1 and 2 (reported previously in Figures 6 and 7). *T* test statistics (reported with their corresponding Cohen's *d* effect size) show that there were no statistically significant differences observed between comparable rates of the two experiments. Note that due to the nature of the comparisons being done, these plots are presented with between-subject confidence intervals which differ from the within-subject confidence intervals presented in Figures 6 and 7.

patient safety and outcomes. Second, although our method can account for additional harmonics, doing so requires more computational time and resources. So pronounced is this, that it has the potential to limit the applicability of the method. Thus, by showing that we only need to account for the primary harmonics in analyses, our results expand the potential usefulness and approachability of our method. This should help enable the use of our method in the analysis of the planned changes to the international standards and by medical device companies designing medical alarms. Third, our results validate the previous findings that have been made using our method. This includes evaluations of the standard's reserved alarm sounds (Bolton et al., 2018) and standard-compliant alarms used in real telemetry monitoring systems (Bolton et al.,

2018). These analyses found compelling problems with these alarms. Thus, the previous results along with the validation presented in this paper suggest that there could be serious masking problems with the alarms of IEC 60601-1-8. Future work should systematically explore when and how masking can manifest in the standard.

Beyond the masking results, our experiments also provide some troubling data about the standard. In particular, across both experiments, the minimum miss and false alarm rates (even in the absence of masking) was approximately 30%. This means that even without masking, the alarm sounds prescribed by the standard can be very difficult to distinguish from the others. Although we used a different experimental design, our results are consistent with research by Lacherez, Seah, and Sanderson (2007), who found that alarm sounds from the standard were very difficult to distinguish from each other when they played concurrently. As such, it is clear that changes will need to be made to the sounds of the alarm standard to make them more distinguishable. The work presented in this paper is being conducted concurrently with a number of other coordinated efforts (Edworthy et al., 2018) to address shortcomings in IEC 60601-1-8 and recommend improvements. Thus, results from the work presented in the paper will be used to help improve the general distinguishability of standard alarm sounds.

As with any study, there were some limitations to our experiments. These and future work are discussed in the following sections.

Additional Experimental Considerations

There are factors that limit the realism of our experiment: We only considered single tones from alarm melodies; experiments were conducted in a quiet controlled laboratory (not a realistic environment); and participants were able to give the experiment their undivided attention (something extremely unlikely in a health care scenario). All of these factors were intentionally chosen to allow the experiment to isolate the effect of masking and minimize the impact of other limits on human perception, attention, and cognition. However, future work could investigate the true impact masking would have on alarm identification in more realistic contexts. Given the strong impact masking had on detection in the ideal listening conditions in our experiment, we would expect even worse detection performance in more realistic settings. Future work should investigate what proportion of alarm perceivability is attributable to simultaneous masking in realistic medical environments.

Experiment 2 only considered one method for including additional frequencies in alarm sounds. Although the parameters for these that were used in our experiment followed common guidelines (Thompson, 2010), it is possible that different parameters could improve alarm distinguishability. In particular, alarms could possibly be made to be more salient by using additional harmonics that are not integer multiples of the primary one, thus creating harmonic dissonance. This should be the subject of future research.

Investigation of Bias

In both experiments, participants had a larger, positive bias in the masking condition than in the nonmasking condition (although this difference was only statistically significant in Experiment 2). This means that they tended to say "Yes" in masking trials more often than in the nonmasking ones. It is not clear why this occurred. One possibility has to do with the fact that in trials with masking, masking sounds could sometimes sound slightly "warbly" (trilling or quavering). This may be caused by physical interactions (called beating; see Levitin, 2006) between the frequencies of the masking and masked sounds. It is possible that this "warbliness" was used by some participants as a cue that the judgment sound was present. This should be investigated in future research.

Additional Alarm Sounds

As part of the larger effort to revise the standard (Edworthy et al., 2018), researchers are designing new alarms that are more complex and harmonically rich than the current melodies of tones, though the melodic patterns will likely remain through legacy support. Thus, while the results presented here will remain topical, the psychoacoustics of simultaneous masking

validated in this work are not appropriate for the new alarm sounds. However, there are other masking curves that use different formulations of spreading functions and Δ than those shown in Equations 4 and 5, respectively, that can represent the masking effect of more complex sounds (Bosi & Goldberg, 2003; You, 2010). Future work should investigate which of these is most appropriate for the new sounds and use experimental validation (like the one presented here) to assess their predictive power.

Additional Application Domains

The focus of the presented research is exclusively on medical alarms. However, alarms are used to alert humans to problems in many other safety critical domains including aviation (Bliss & Acton, 2003), industrial control rooms (Rothenberg, 2009), and driving (Bliss, 2003). Many of the same problems that affect medical alarms can also manifest in these other areas. In fact, there have been a few instances of design recommendations for avoiding the effects of masking in these industries (Begault, Godfroy, Sandor, & Holden, 2007; Patterson, 1982; Patterson & Mayfield, 1990; Wolfman, Miller, & Volanth, 1996). However, to the best of our knowledge, nobody has investigated whether simultaneous masking does in fact manifest in these environments. Thus, future work should determine whether simultaneous masking is occurring and, if so, how our methods could be used to assess its potential risks.

ACKNOWLEDGMENTS

The research reported in this paper was supported by the Agency for Healthcare Research and Quality under award number R18HS024679. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University at Buffalo. Informed consent was obtained from each participant.

KEY POINTS

- The alarms prescribed by the IEC 60601-1-8 international standard are theoretically susceptible to simultaneous masking.
- This work validates that the psychoacoustics of simultaneous masking can accurately predict the perceivability of standard-compliant medical alarm sounds using two signal detection experiments.
- The experiments showed that the psychoacoustics did accurately predict the perceivability of alarm sounds based on whether their primary harmonics were masked.
- The results further validate that a formal methods model developed in previous work can accurately predict whether humans will hear IEC 60601-1-8-compliant alarms.
- The results will influence methods for detecting masking in medical alarm designs as well as updates to the international standard.

REFERENCES

Ambikairajah, E., Davis, A., & Wong, W. (1997). Auditory masking and MPEG-1 audio compression. *Electronics & Communication Engineering Journal*, *9*(4), 165–175.

Baumgarte, F., Ferekidis, C., & Fuchs, H. (1995). A nonlinear psychoacoustic model applied to ISO MPEG layer 3 coder. *Proceedings of the Audio Engineering Society Convention*. New York, NY: Audio Engineering Society. Retrieved from https://pdfs.semanticscholar.org/b929/0a2eac6b28e4ed169bfcc637ceeb0e0ae50d.pdf

Begault, D. R., Godfroy, M., Sandor, A., & Holden, K. (2007). Auditory alarm design for NASA CEV applications. *Proceedings of the 13th International Conference on Auditory Display* (pp. 131–138). Montreal, Québec, Canada.

Bliss, J. P. (2003). Investigation of alarm-related accidents and incidents in aviation. *The International Journal of Aviation Psychology*, *13*, 249–268.

Bliss, J. P., & Acton, S. A. (2003). Alarm mistrust in automobiles: How collision alarm reliability affects driving. *Applied Ergonomics*, *34*, 499–509.

Bolton, M. L., Edworthy, J. R., & Boyd, A. D. (2018). A formal analysis of masking between reserved alarm sounds of the IEC 60601-1-8 international medical alarm standard. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*, 523–527. doi:10.1177/1541931218621119

Bolton, M. L., Edworthy, J. R., Boyd, A. D., Wei, J., & Zheng, X. (2018). A computationally efficient formal method for discovering simultaneous masking in medical alarms. *Applied Acoustics*, *141*, 403–415.

Bolton, M. L., Hasanain, B., Boyd, A. D., & Edworthy, J. R. (2016). Using model checking to detect masking in IEC 60601-1-8-compliant alarm configurations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*, 636–640. doi:10.1177/1541931213601146

Bosi, M., & Goldberg, R. E. (2003). *Introduction to digital audio coding and standards*. New York, NY: Springer.

Brandenburg, K., & Bosi, M. (1997). Overview of MPEG audio: Current and future standards for low bit-rate audio coding. *Journal of the Audio Engineering Society, 45*(1/2), 4–21.

Brandenburg, K., & Stoll, G. (1994). ISO/MPEG-1 audio: A generic standard for coding of high-quality digital audio. *Journal of the Audio Engineering Society*, *42*, 780–792.

Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). *Model checking*. Cambridge, MA: MIT Press.

Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: A comment on Baguley (2012). *Behavior Research Methods*, *46*, 1149–1151.

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, *121*, 275–277.

ECRI Institute. (2014, November). Top 10 health technology hazards for 2015. *Health Devices*. Retrieved from https://www.ecri.org/Resources/Whitepapers_and_reports/Top_Ten_Technology_Hazards_2015.pdf

ECRI Institute. (2018, October). 2019 top 10 health technology hazards. *Health Devices*. Retrieved from https://www.ecri.org/Resources/Whitepapers_and_reports/Haz_19.pdf

Edworthy, J. (2013). Medical audible alarms: A review. *Journal of the American Medical Informatics Association*, *20*, 584–589.

Edworthy, J., & Hellier, E. (2005). Fewer but better auditory alarms will improve patient safety. *Quality and Safety in Health Care*, *14*, 212–215.

Edworthy, J., & Hellier, E. (2006). Alarms and human behaviour: Implications for medical alarms. *British Journal of Anaesthesia*, *97*(1), 12–17.

Edworthy, J., McNeer, R. R., Bennett, C. L., Dudaryk, R., McDougall, S. J., Schlesinger, J. J., . . . Osborn, D. (2018). Getting better hospital alarm sounds into a global standard. *Ergonomics in Design*, *26*(4), 4–13.

Edworthy, J., & Meredith, C. S. (1994). Cognitive psychology and the design of alarm sounds. *Medical Engineering & Physics*, *16*, 445–449.

Fastl, H., & Zwicker, E. (2006). *Psychoacoustics: Facts and models* (*Vol. 22*). New York, NY: Springer.

Green, D. M. (1967). Additivity of masking. *The Journal of the Acoustical Society of America*, *41*, 1517–1525.

Hasanain, B., Boyd, A., & Bolton, M. L. (2014). An approach to model checking the perceptual interactions of medical alarms. *Proceedings of the International Annual Meeting of the Human Factors and Ergonomics*, *58*, 822–826. doi:10.1177/1541931214581173

Hasanain, B., Boyd, A., & Bolton, M. L. (2016). Using model checking to detect simultaneous masking in medical alarms. *IEEE Transactions on Human-machine Systems*, *46*, 174–185.

Hasanain, B., Boyd, A. D., Edworthy, J., & Bolton, M. L. (2017). A formal approach to discovering simultaneous additive masking between auditory medical alarms. *Applied Ergonomics*, *58*, 500–514.

Humes, L. E., & Jesteadt, W. (1989). Models of the additivity of masking. *The Journal of the Acoustical Society of America*, *85*, 1285–1294.

IEC 60601-1-8:2006/AMD1:2012. (2012). *Medical electrical equipment—Part 1-8: General requirements for basic safety and essential performance—Collateral standard: General requirements, tests and guidance for alarm systems in medical electrical equipment and medical electrical systems*. Geneva, Switzerland: International Electrotechnical Commission.

The Joint Commission. (2013a, April). Medical device alarm safety in hospitals. *Sentinel Event Alert*, *50*, 1–3.

The Joint Commission. (2013b, July). NPSG.06.01.01: Improve the safety of clinical alarm systems. *Joint Commission Perspectives*, 33, 1–4.

Konkani, A., Oakley, B., & Bauld, T. J. (2012). Reducing hospital noise: A review of medical device alarm management. *Biomedical Instrumentation & Technology*, *46*, 478–487.

Lacherez, P., Seah, E., & Sanderson, P. (2007). Overlapping melodic alarms are almost indiscriminable. *Human Factors*, *49*, 637–645.

Levitin, D. J. (2006). *This is your brain on music: The science of a human obsession*. New York, NY: Penguin.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*, 401–413.

McNicol, D. (2005). *A primer of signal detection theory*. Mahwah, NJ: Lawrence Erlbaum.

Meredith, C., & Edworthy, J. R. (1995). Are there too many alarms in the intensive care unit? An overview of the problems. *Journal of Advanced Nursing*, *21*, 15–20.

Momtahan, K., Hetu, R., & Tansley, B. (1993). Audibility and identification of auditory alarms in the operating room and intensive care unit. *Ergonomics*, *36*, 1159–1176.

Patterson, R. D. (1982). *Guidelines for auditory warning systems on civil aircraft*. London, England: Civil Aviation Authority.

Patterson, R. D., & Mayfield, T. F. (1990). Auditory warning sounds in the work environment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *327*, 485–492.

Pennsylvania Patient Safety Authority. (2009). Connecting remote cardiac monitoring issues with care areas. *Pennsylvania Patient Safety Advisory*, *6*(3), 79–83. Retrieved from http://patientsafety.pa.gov/ADVISORIES/Pages/200909_79.aspx

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*, 125–126.

Rothenberg, D. H. (2009). *Alarm management for process control: A best-practice guide for design, implementation, and use of industrial alarm systems*. New York, NY: Momentum Press.

Schroeder, M. R., Atal, B. S., & Hall, J. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, *66*, 1647–1652.

See, J. E., Warm, J. S., Dember, W. N., & Howe, S. R. (1997). Vigilance and signal detection theory: An empirical evaluation of five measures of response bias. *Human Factors*, *39*, 14–29.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

Stead, W. W., & Lin, H. S. (Eds.). (2009). *Computational technology for effective health care: Immediate steps and strategic directions*. Atlanta, GA: National Academies Press.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, *1*, 155–182.

Thompson, C. (2010). *ISO/IEC 60601-1-8, Patterson and other alarms in medical equipment sample alarm sounds—Sirens, buzzers and other sounds*. Retrieved from http://www.anaesthesia.med.usyd.edu.au/resources/alarms/

Toor, O., Ryan, T., & Richard, M. (2008). Auditory masking potential of common operating room sounds: A psychoacoustic analysis. In *Anesthesiology* (Vol. 109, p. A1207). Park Ridge, IL: American Society of Anesthesiologists.

Vasseur, D. A., & Yodzis, P. (2004). The color of environmental noise. *Ecology*, *85*, 1146–1152.

Vockley, M. (2014). *Clinical alarm management compendium*. Arlington, VA: AAMI Foundation.

Wolfman, G. J., Miller, D. L., & Volanth, A. J. (1996). An application of auditory alarm research in the design of warning sounds for an integrated tower air traffic control computer system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *40*, 1002–1006. doi:10.1177/154193129604001910

You, Y. (2010). *Audio coding: Theory and applications*. New York, NY: Springer Science + Business Media.
Zwicker, E., & Feldtkeller, R. (1967). *Das ohr als nachrichten-empfänger* [The ear as a communication receiver]. Stuttgart, Germany: Hirzel Verlag.

Matthew L. Bolton is an associate professor of industrial and system engineering at the University at Buffalo, The State University of New York. He received the PhD in systems engineering in 2010 from the University of Virginia, Charlottesville, USA.

Xi Zheng is a PhD student in industrial and systems engineering at the University at Buffalo, The State University of New York. She received the BS in electronic commerce in 2011 from Southwest University, Chongqing, China.

Meng Li is a human factors engineer and UX designer intern at Medtronic. He received the PhD in industrial and systems engineering in 2018 from the University at Buffalo, The State University of New York, USA.

Judy Reed Edworthy is the director of the Cognition Institute and a professor of applied psychology at the University of Plymouth. She received the PhD in experimental psychology in 1984 from the University of Warwick, UK.

Andrew D. Boyd is an associate professor of biomedical and health information sciences at the University of Illinois Chicago. He received the MD in 2002 from the University of Texas Southwestern Medical School, Dallas, USA.

Date received: February 23, 2019
Date accepted: June 14, 2019

Classification of Attentional Tunneling Through Behavioral Indices

Sean W. Kortschot and **Greg A. Jamieson**, University of Toronto, Ontario, Canada

Objective: The objective of this study was to develop a machine learning classifier to infer attentional tunneling through behavioral indices. This research serves as a proof of concept for a method for inferring operator state to trigger adaptations to user interfaces.

Background: Adaptive user interfaces adapt their information content or configuration to changes in operating context. Operator attentional states represent a promising class of triggers for these adaptations. Behavioral indices may be a viable alternative to physiological correlates for triggering interface adaptations based on attentional state.

Method: A visual search task sought to induce attentional tunneling in participants. We analyzed user interaction under tunnel and non-tunnel conditions to determine whether the paradigm was successful. We then examined the performance trade-offs stemming from attentional tunnels. Finally, we developed a machine learning classifier to identify patterns of interaction characteristics associated with attentional tunnels.

Results: The experimental paradigm successfully induced attentional tunnels. Attentional tunnels were shown to improve performance when information appeared within them, but to hinder performance when it appeared outside. Participants were found to be more tunneled in their second tunnel trial relative to their first. Our classifier achieved a classification accuracy similar to comparable studies (area under curve = 0.74).

Conclusion: Behavioral indices can be used to infer attentional tunneling. There is a performance trade-off from attentional tunneling, suggesting the opportunity for adaptive systems.

Application: This research applies to adaptive automation aimed at managing operator attention in information-dense work domains.

Keywords: attentional processes, adaptive automation, attentional tunneling, passive data monitoring, machine learning

Address correspondence to Sean W. Kortschot, University of Toronto, 5 King's College Road, Toronto, Ontario M5S 3G8, Canada; e-mail: sean.kortschot@mail.utoronto.ca.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 973–986
DOI: 10.1177/0018720819857266
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Decision support systems (DSSs) have become commonplace in modern control rooms (Dongen & Maanen, 2013). The principle role of the DSS is to aid operators in effectively extracting meaningful insights from the often copious information that is available to them. DSSs thus serve as attentional guidance tools by highlighting important or anomalous behavior in the information space of interest (Corchado & Herrero, 2011). In spite of advances to DSSs, the attentional load imposed on operators in many domains remains significant. This presents an opportunity for *adaptive user interfaces*, which adapt their information content, configuration, or interactions to facilitate operator performance and workload (Feigh, Dorneich, & Hayes, 2012).

Adaptations can be *triggered* by either spatiotemporal changes or changes in the system, environment, task, or operator (Feigh et al., 2012). We focus here on the *operator measurement* class of triggers, which respond to an operator's mental or physical state. Feigh et al. (2012) define two methods for measuring an operator's state: physiology and performance. Physiological state inference relies on technologies such as electroencephalograms (EEG) or electrocardiograms (ECG), an example being the *Communication Scheduler*, which uses EEG and ECG to infer soldiers' workload to modulate communications (Dorneich, Whitlow, Mathan, Carciofini, & Ververs, 2005). Performance state inference relies on real-time measures of operator performance assuming an inverse proportionality to workload (Feigh et al., 2012). A central challenge limiting the implementation of performance-based operator state inference, however, is that most task domains lack an accurate and continuous performance metric. This lack, coupled with increased accessibility and

accuracy (McLane et al., 2015; Verwey, Shea, & Wright, 2015) as well as decreasing invasiveness of apparatus (Mukhopadhyay & Lay-Ekuakille, 2010), has led most research on operator state inference to focus on physiological measures. Some notable examples of physiological state inference include detection of mind wandering (Baldwin et al., 2017), task engagement (Berka et al., 2007), mental workload (Borghetti, Giametta, & Rusnock, 2017; Wilson & Russell, 2004), and various emotional states (Lee & Hsieh, 2014). Although these studies demonstrate the promise of physiological state inference, practical applications still require initial investment and users willing to use the necessary recording devices and share their biometric data.

A third approach for operator measurement, not described in Feigh et al. (2012), is passive data monitoring (PDM), which focuses on *behavior* rather than *performance* by examining streams of interaction data that are inherent to a task (Kortschot et al., 2018; Palmius et al., 2016). PDM exploits data that exist irrespective of performance (e.g., cell phone location from GPS, cursor position in desktop applications, etc.) and examines it for patterns that align with cognitive events or states. PDM is not intended to replace physiological inference in all domains. Instead, it represents a viable alternative in domains with behavioral data that is both rich and readily available and where outfitting operators with biometric recording devices is impractical, unfeasible, or unnecessary. PDM sacrifices some of the precision that biometrics can offer (e.g., cursor tracking is a less accurate measure of attentional focus than eye tracking) for significantly increased practicality. The efficacy of PDM for inferring cognitive events or states has been demonstrated for attentional switching (Kortschot et al., 2018), depression onset detection (Palmius et al., 2016), information overload (Mac Aoidh, Bertolotto, & Wilson, 2012), and driver drowsiness detection (McDonald, Lee, Schwarz, & Brown, 2014).

Both PDM and physiological state inference share the fundamental challenge of labeling what data belong to what cognitive state. Palmius et al. (2016) labeled impending depressive episodes via weekly questionnaires. Although this method is well suited to enduring cognitive states, it is incompatible with more transient states that adaptive automation typically targets. Early work in physiological state inference solved the labeling issue by inducing a state and examining how peoples' physiology responded (Kramer, 1991). A similar approach can be used for PDM.

The attentional state that we focus on in the present research is *attentional tunneling*, which Wickens and Alexander (2009) operationalized as

> the allocation of attention to a particular channel of information, diagnostic hypothesis, or task goal, for a duration that is longer than optimal, given the expected cost of neglecting events on other channels, failing to consider other hypotheses, or failing to perform other tasks. (p. 182)

We operationalize attentional tunneling slightly differently, describing it in relation to the *potential* cost of neglecting events rather than the *expected* cost, as in many instances attentional tunneling can be conflated with directed attention, which can serve as an adaptive mechanism (Dixon et al., 2013). Attentional tunneling is believed to be a significant contributor to the accident at Three Mile Island (Rubenstein & Mason, 1979), most aviation accidents involving controlled flight into terrain (Shappell & Wiegmann, 2003), and many other incidents.

Several factors can precipitate attentional tunnels. These include increased cognitive load (Rantanen & Goldberg, 1999; Williams, 1995), display design (Wickens & Alexander, 2009), conversation (Atchley & Dressel, 2004), and operator fatigue (Mills, Spruill, Kanne, Parkman, & Zhang, 2001). Rogé, Kielbasa, and Muzet (2002) demonstrated that task priority and complexity can lead to attentional tunnels, with operators overly fixating on tasks that they perceive to be the most important. Many of these factors are prevalent in information dense spaces. Therefore, attentional tunneling represents a promising candidate attentional state for treatment through DSS design.

This article presents an experiment aimed at inducing attentional tunneling and developing a

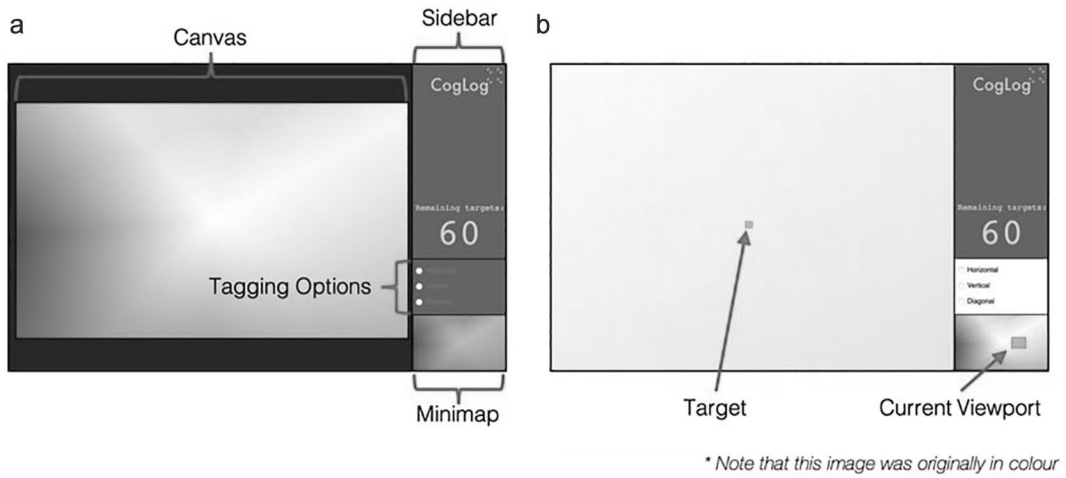


Figure 1. CogLog platform used for experimentation. (a) Example of CogLog when fully zoomed-out. (b) Example of CogLog when fully zoomed-in on the target and when the target has been clicked on, which activated the tagging options. All red brackets, arrows, and labels are included for the purpose of this figure. Note that the target in Figure 1b has been darkened for the purpose of this figure.

machine learning classifier based on PDM that can detect when users are in that state. We examine the viability of this method in a low-fidelity environment and describe the performance trade-offs stemming from attentional tunnels. We hypothesize that the induction of an attentional tunnel will improve performance when information appears within that tunnel at the expense of degrading performance when information appears outside of it. Finding this performance trade-off will verify successful induction of an attentional tunnel and therefore validate that the resulting machine learning classifier is identifying attentional tunnels. This research serves as a proof of concept for PDM.

METHODS

Experimental Platform

This study was conducted on a novel experimental platform called the Cognitive Logger (CogLog). CogLog is an online platform used to collect behavioral datasets for developing machine learning classifiers. The version of CogLog used in the present study (CogLog_AT.3) consisted of two main elements: a canvas and a sidebar (see Figure 1; Supplementary Material S1). The canvas was 1,000 × 600 pixels and its background was set to a continuous color wheel covering the full visible color spectrum. Participants were able to *zoom in* on different areas of the canvas, which caused a smaller portion of the canvas to occupy a greater portion of the viewport (i.e., the portion of the canvas that was currently in view). They could also *pan across* the canvas by clicking and dragging, which caused areas of the canvas that were adjacent to the participants' viewport to be brought into view. By doing this, participants were able to search for the presence of targets that appeared, one at a time, in random locations on the canvas.

Targets were small (3 × 3 pixel) striped patches containing either horizontal, vertical, or diagonal lines. They had an opacity of 12% with a *multiply mix-blend-mode*, which caused their RGB values to be multiplied by the RGB values of the section of the canvas on which they appeared (Budd & Björklund, 2016). For example, if a target appeared in a red area of the canvas, the red values of the target would be increased disproportionately to the green and blue values, thereby causing the target to render with a red tint. This allowed for approximate equivalence of detection difficulty regardless of the background color on which the target appeared. The size and transparency of targets ensured that participants would have to zoom in

and scan subsections of the canvas for successful detection and classification.

The sidebar consisted of a counter, tagging options, and a minimap. The counter indicated how many targets remained until the end of the experiment. Clicking on a target activated the tagging options (see Figure 1b), which were radio buttons that allowed participants to classify the target as being horizontal, vertical, or diagonal. The minimap allowed participants to see where their current viewport fell within the overall canvas with a red, translucent square (see Figure 1b).

CogLog recorded all user interactions, which included cursor position relative to both the screen and canvas, the location and size of the canvas relative to the participant's screen, and the classifications made by the participant. From these data, we were able to derive all zooming and panning behaviors. CogLog sampled user data every time they executed an action (i.e., moved their cursor or moved the canvas). This was done rather than sampling at fixed rates so that we could later transform the data to sample at any rate less than the maximum sampling rate of the logger.

CogLog is a *behavioral analog* to a cybersecurity microworld simulation (see Kortschot et al., 2017), meaning that the input features (i.e., scroll to zoom, click and drag to pan) are identical between CogLog and the simulation. Therefore, demonstrating that attentional states can be classified using the suite of interactions in this environment will suggest the viability of the approach in higher fidelity environments.

Experimental Design

The experiment employed a single factor within-subjects design. Participants each completed four trials, two in the *tunnel* condition and two in the *non-tunnel* condition. There were two classes of targets used in both conditions: *prime* targets, which were all targets in each trial except for the last target, and a *test* target, the final target in each trial. Targets appeared one at a time with the next target generated upon the classification of the current target. We labeled the area in which targets could appear the *active area*, the boundaries of which remained unmarked to participants. In the tunnel condition, the active area was a subsection of the canvas approximately one tenth the size of the total canvas area. The active area was randomly positioned at the start of each trial and all prime targets appeared randomly within this region. In the non-tunnel condition, the active area for prime targets occupied the entire canvas. The prime phase of each trial was intended to *prime* an attentional state in the participant (e.g., Friedman, Fishbach, Förster, & Werth, 2003). The test target in both tunnel and non-tunnel conditions could appear anywhere on the canvas, provided that it was at least 300 × 200 pixels away from the final prime target. Figure 2 illustrates the experimental conditions.

Both the color wheel that was used as the canvas background and the minimap were intended to implicitly remind participants where they were located in the canvas. Essentially, by having all prime targets appear in a small box in the tunnel condition, they also appeared in the same color space. This was expected to bias participants to a particular region and color such that navigating to a new area felt foreign during the test phase.

During the prime phase, participants were never given any information about the distribution governing target generation. Therefore, if they behaved differently between conditions it could be only attributed to them updating their expectations about where future targets were likely to appear given where previous targets had appeared before. Furthermore, because there was no difference in how the test target was generated between conditions, any difference in participant behavior during the search for the test target can solely be attributed to the distribution of the prime targets that preceded it. Critically, this paradigm ensures that we are not distinguishing between behavior from two different tasks, but rather between behavior from the same task performed under two different attentional states.

The majority of the collected data came from the prime phase of the study. This meant that there was insufficient data from the test phase to develop a reliable machine learning classifier. We therefore used traditional statistics to examine any differences in behavior during the test

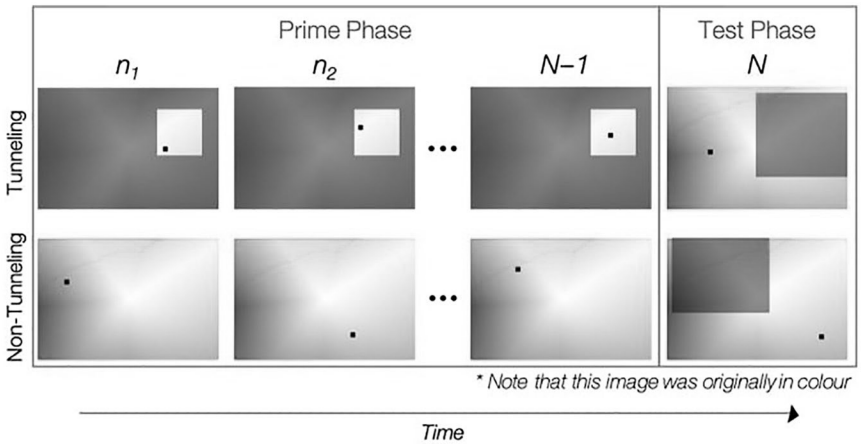


Figure 2. Top-row: The tunneling condition. The active area (unshaded) was approximately 10% of the canvas size. All prime targets appeared within this area. Bottom-row: The non-tunneling condition. The active area (unshaded) occupied the entire canvas for all prime targets. In both conditions, the test target could appear anywhere on the canvas outside of a 300 × 200 pixel box from the final prime target ($N-1$). For both conditions, N ranged from 13 to 16 at random. Targets here are darker and larger for illustrative purposes.

phase, which identified whether participants succumbed to attentional tunnels in the tunnel condition relative to the non-tunnel condition. Our classifier was then trained on the prime phase data, and sought to determine a pattern of behavior that was indicative of a developing attentional tunnel.

Procedure

Experimentation was completed on Amazon's Mechanical Turk. Sixty participants were paid a flat rate of Can$10.00, thereby incentivizing them to complete the task as fast as possible. Due to a database error, only data from 50 participants was used in our analysis. Following a description of the study's requirements, participants were given a screening questionnaire that presented sample targets on different areas of the canvas's color wheel. This was designed to test screen resolution, visual acuity, and color vision and ensured that anyone who advanced to experimentation would be able to complete the task. Participants were then given an interactive training regimen that detailed how to interact with the interface, presented videos of an expert user performing the task, and presented participants with five practice targets to search for. Participants could only advance to experimentation once they had successfully classified the five practice targets.

Each participant completed four trials and classified a total of 58 targets. The number of targets within each trial ranged from 13 to 16 in random order and the sequencing of tunnel and non-tunnel conditions was fully randomized. Although participants were never informed of the experimental condition that they were in and never saw the bounding box for the active area, astute participants may have inferred the two different distribution algorithms used to generate targets. They may have therefore realized that the final target in their previously experienced tunnel trial fell outside of the prime phase's active area. The counter (see Figure 1) therefore began at 60 rather than 58 so that it never informed participants they were on the last target of the experiment.

All research herein complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Toronto. Informed consent was obtained from each participant.

RESULTS

Participants accurately classified 99.07% of possible targets. Therefore, all results will focus on the temporal aspects of performance rather than accuracy.

Data Cleaning

The data were transformed from the input-dependent method to a fixed sampling rate of one sample per 75 ms. This forward filled stretches with no interactions and removed observations during stretches where the user was interacting at a rate higher than once every 75 ms. For all statistical tests and for our machine learning classifier, we excluded all behavior recorded during the search for the first target because the time it took for a participant to identify this target was an artifact of the participants' search relative to the random position of the target. We also excluded behavior recorded during search for the second target because a tunnel could not have been reinforced after a participant only identified the location of one target. Because this reduced the number of targets in tunnel trials, our performance metrics depict average search time per target rather than the completion time of the overall trials.

Characterization of Attentional Tunneling

To characterize whether the paradigm successfully induced attentional tunnels, we needed to examine several aspects of behavior during both prime and test phases. Unfortunately, no perfect measure exists for what behavior represented an attentional tunnel. This is partially due to the remote nature of the experiment but also because of the relative novelty of studying attentional tunneling purely through behavior. Therefore, we needed to look at several proxy measures and determine whether, when taken together, they suggest an effect.

Our principal metric for characterizing attentional tunneling was the time that it took to classify the test target. The hypothesis driving this analysis is that if participants developed an attentional tunnel during the prime phase of tunnel trials, they should expect the test target to appear within that tunnel and therefore continue their search in that area, which should result in longer classification times for the test target. The experimental paradigm likely induced different levels of attentional tunneling during the prime phase depending on the trial and the participant. The degree to which an attentional tunnel was induced was expected to influence participants' expectations about where the test target would appear. We therefore used *performance* and *degree of interaction* during the prime phase of each trial as proxy measures for the degree of attentional tunneling that was induced in that particular trial. We measured performance in the prime phase of each trial as the amount of time that it took on average to classify a target within that trial. Degree of interaction was a measure of the net amount of cursor, pan, and zoom movement, all normalized between zero and one, per target during the prime phase of each trial. We performed generalized linear mixed models specifying participant, trial number, prime phase performance, and degree of interaction as random variables. We found that it took participants significantly longer to find the test target in the tunnel condition (M = 16.48 s, SE = 0.09) compared with the non-tunnel condition (M = 13.36 s, SE = 0.10); $F(1, 148) = 4.55$, $p = .03$.

Our second measure for characterizing attentional tunnels was *test target search bias*, which is a measure of whether participants' search during the test phase was biased toward one area of the canvas. We calculated this in tunnel trials by taking the moment in the test phase that no part of the preceding prime phase's active area remained within the participants' viewports (see Figure 2). In non-tunnel trials, we calculated this by taking the moment in the test phase that no part of that test phase's inactive area remained within participants' viewports. Participants' test target search bias was significantly shorter in the non-tunnel condition (M = 5.54 s, SE = 0.09) compared with the tunnel condition (M = 7.12 s, SE = 0.08); $F(1, 129) = 10.47$, $p = .002$.

To further test whether the experimental design successfully induced attentional tunnels, we examined three additional proxies. First, we examined the average classification time per

target between tunnel and non-tunnel conditions to determine whether participants updated their beliefs about future target generation. We found that it took participants less time on average to find an individual target in the tunnel condition (M = 14.62 s, SE = 0.08) compared with the non-tunnel condition (M = 22.47 s, SE = 0.08); $F(1, 152) = 87.74$, $p < .0001$. Second, we examined the mean zoom level during search for the test target to determine the average amount of the canvas that participants were searching at each point. We found a significant effect, with participants viewing a smaller portion of the canvas per logged action in the tunnel condition (M = 18%, SE = 0.07) compared with the non-tunnel condition (M = 21%, SE = 0.07); $F(1, 135691) = 583.01$, $p < .0001$. Finally, we examined whether detection time of an individual target and degree of interaction within that trial's prime phase were influenced by whether it was the first or second tunnel trial. We found significant effects for both, showing that participants found individual targets faster in the prime phase of their second tunnel trial (M = 10.72 s, SE = 0.07) while also interacting less (M = 0.58, SE = 0.10) compared with their first (M = 13.63 s, SE = 0.07; M = 0.76, SE = 0.11); $F(1, 49) = 10.72$, $p = .002$; $F(1, 49) = 4.15$, $p = .047$. However, there was no difference in the amount of time that it took for participants to find the test target between their first (M = 20.54 s, SE = 0.09) and second trials (M = 19.54 s, SE = 0.09); $F(1, 48) = 0.01$, $p = .91$.

DISCUSSION OF BEHAVIORAL FINDINGS

Taken together, the results from the first phase of our analysis indicate that the experimental paradigm successfully induced attentional tunnels in the tunnel condition relative to the non-tunnel condition. Although none of the measures that we included are perfect representations of attentional tunneling, the fact that they nearly all point in the same direction and that they tend to align with the behavioral aspects of the Wickens and Alexander (2009) definition of attentional tunnels suggests that the intended effect was achieved.

We found that participants searched the area around the final prime target significantly longer in the tunnel condition, indicating that they were overly focused on the area where they believed the next target most likely to appear and that this focus persisted in the absence of new targets. This additional focus delayed their target detection time when searching for the test target. This result, coupled with the finding that participants found targets within the prime phase of the tunnel condition significantly faster than in the prime phase of the non-tunnel condition, demonstrates a clear performance trade-off: Improved performance when information appeared within an attentional tunnel at the cost of poorer performance when information appears elsewhere. This supports our operationalization's focus on the potential rather than the expected cost of an attentional tunnel. The degree to which this trade-off is considered in the design of DSSs should be dependent on the domain of interest. For example, in a safety critical system like a nuclear power plant, it may be advisable to deter attentional tunnels even if they improve performance the majority of the time as, in rare edge cases, they may allow dangerous operating conditions to develop.

Interestingly, participants exhibited more tunneled behavior during the prime phase of their second tunnel trial compared with their first, suggesting a relationship between experience with a system and susceptibility to attentional tunnels. The stronger tunnels in the prime phase of the second tunnel trial resulted in improved performance, which did not carry over to the test phase of that trial. This may suggest that the presumed performance gains were offset by stronger preceding tunnels. The degree to which these effects carry over to higher fidelity simulations and over longer time spans where operators gain more experience with a system needs to be evaluated. However, the alignment of these results with the literature (e.g., Briggs, Hole, & Turner, 2018; Dixon et al., 2013; Wickens & Alexander, 2009) indicates that they are likely to translate.

MACHINE LEARNING CLASSIFICATION

The findings from our initial analyses suggest that the experimental paradigm was successful in inducing attentional tunnels in the tunnel condition relative to the non-tunnel

condition. We will now present a machine learning classifier trained on data from the prime phase of the experimental paradigm. The goal of this classifier was to identify patterns of behavior that were characteristic of an attentionally tunneled state relative to a non-tunneled state.

Unlike many other classification tasks (e.g., image classification), our paradigm does not assume perfect ground truth labeling. Although we determined that the experimental paradigm successfully induced attentional tunnels in the tunnel condition overall, it was likely that there were some trials where it failed (see Li et al., 2015, for a related approach). Therefore, to ensure that the data that we were passing into the classifier were reflective of its label, we examined both the degree of interaction and the time required to identify a target within the prime phase of both tunnel and non-tunnel trials. These measures determined whether participants updated their search strategy in tunnel trials relative to non-tunnel trials. If they interacted with the platform more, or took longer per target in a tunnel trial compared with their averages in non-tunnel trials, then it was determined that trial failed to induce the desired attentional tunnel. We identified 13 such trials out of a total of 208 trials and did not include their data in our classifier.

Inference Techniques

Our data were multivariate time series data and consisted of the X/Y coordinates of the user's cursor relative to the canvas, the zoom level of the viewport, and the X/Y coordinates of the canvas itself. Using these measures we calculated 12 total features, which included the speed of the cursor and screen as well as binned features that essentially reduced the canvas to a 20 by 20 grid, thereby decreasing the resolution of the data to cancel out some noise. At no point was the location of the targets passed into the classifier. Therefore, all classification was based purely on how the participant interacted with the platform and hinged on participants updating their search behavior as they received increasing evidence about the distribution of targets.

Long short-term memory (LSTM) recurrent neural networks are well suited to this type of classification task as they can represent past information in their internal states and use these states to process new information, but without suffering from the vanishing gradient problem that can limit traditional recurrent neural networks (Hochreiter & Schmidhuber, 1997; Lipton, Berkowitz, & Elkan, 2015). LSTMs control how internal states update and output at each time step via gating functions (Hochreiter & Schmidhuber, 1997), which allows them to maintain long-term storage of internal states and therefore to exploit distant temporal dependencies within the data (Greff, Srivastava, Koutník, Steunebrink, & Schmidhuber, 2017; Prieto, Alonso-González, & Rodríguez, 2015; Zhao, Chen, Wu, Chen, & Liu, 2017). Furthermore, LSTM networks can be stacked under initial 1-Dimensional convolutional layers (CNN-LSTM) to improve their processing speed and their feature extraction capabilities (Karim, Majumdar, Darabi, & Harford, 2018; Pak, Kim, Ryu, Sok, & Pak, 2018; Tan et al., 2018).

Neural Network Architecture

We used an eight-layer CNN-LSTM to classify the dataset (see Table 1 for model summary). Layers 1 to 4 were convolutional layers and Layers 5 to 7 were LSTM layers composed, respectively, of 256, 128, and 64 neurons with dropout rates of 0.5. The final layer was a fully connected Dense layer with one neuron. The activation level of this neuron represented the classification. The more polarized the activation, the greater confidence that the model had in its classification.

CNN-LSTM models require data to take the form of a 3D array, wherein the first dimension represents the number of sequences in the dataset, the second dimension represents the number of time steps within each sequence, and the final dimension represents the number of features. Our data took the shape (971, 250, 12). The classifier was trained via 10-fold grid search cross validation (GridSearchCV; Pedregosa et al., 2011) using Keras with a TensorFlow backend. Each time window consisted of 250 time steps spaced 75 ms apart thereby covering 18.75 s.

Data Preprocessing

The active area in the tunnel condition was located in a random position on the canvas and

TABLE 1: Model Summary for the CNN-LSTM

Layer	Type	Output Shape	Filters	Kernel Size	Dropout	Recurrent Dropout
0	Input	250 x 12	–	–	–	–
1	1D Conv.	247 x 32	32	4	–	–
2	1D Maxpool	24 x 32	10	1	–	–
3	1D Conv	2 x 32	32	4	–	–
4	1D Maxpool	2 x 128	10	1	–	–
5	LSTM	2 x 256	–	–	0.5	0.35
6	LSTM	2 x 128	–	–	0.5	0.35
7	LSTM	64	–	–	0.5	0.35
8	Dense	1	–	–	–	–
Total parameters		548,097				
Trainable parameters		548,097				

Note. CNN = Convolutional Neural Network; LSTM = long short-term memory.

all targets appeared in random locations within this region. In the non-tunnel condition, targets appeared randomly anywhere on the canvas. Therefore, from trial to trial, even within conditions, behavior was expected to differ significantly. Furthermore, the direction of successive targets within conditions changed at random. This meant that we were unable to simply pass sequence data (i.e., $t_i - t_{i-1}$) into the classifier. For example, in classifying a data stream from an accelerometer of a user walking up a staircase (e.g., Human Activity Recognition dataset; Anguita, Ghio, Oneto, Parra, & Reyes-Ortiz, 2013), researchers could expect that person, regardless of who the person is or the specifics of that staircase, to be traveling in an upward direction (Arif & Kattan, 2015; Vrigkas, Nikou, & Kakadiaris, 2015). Our case is more similar to classifying the activity of playing soccer: The actions should share similar characteristics, but the trajectories of the sequences will differ depending on where the ball is. Figure 3 illustrates how participant behavior can share general properties but differ depending on the specific locations of the targets.

To account for this, we calculated how far participants moved from the first recorded point within each sequence passed to the network, irrespective of direction or location on the canvas. This was done independently for each of the three original features described earlier. This type of preprocessing makes the classifier resilient to the specific location of attentional tunnels.

Because users completed the tunnel condition significantly faster than non-tunnel, we had a roughly 2:1 imbalance in data. We therefore down-sampled the non-tunneling data, removing about half of the sequences. We also normalized all data between −1 and 1. Finally, we removed the 13 trials that failed to induce attentional tunnels according to our degree of interaction and performance criteria (see Data Cleaning section).

CLASSIFIER RESULTS

The principal metric for our classifier was mean area under curve (AUC) across the 10-folds. This is widely recognized as being the best estimator for the performance of a model as it represents an aggregate classification across all classification thresholds (Mason & Graham, 2002). The model achieved a mean AUC of 0.74 across the 10-folds (see Figure 4). The test dataset was composed of 53% tunneling observations, which represents chance performance.

CLASSIFIER DISCUSSION

We demonstrated that classification of an attentional state is possible through behavioral metrics, even with a relatively small, noisy, and

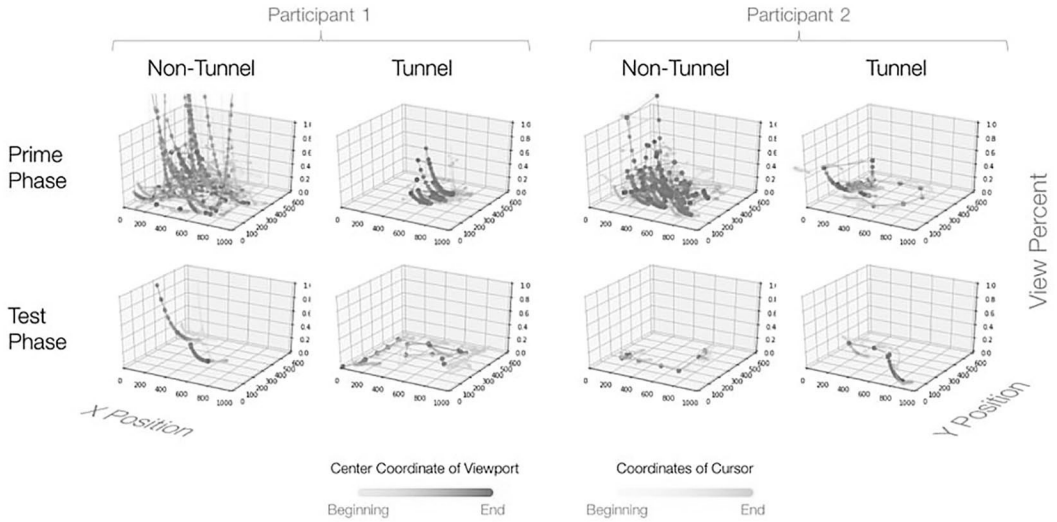


Figure 3. Viewport location and cursor position data from two participants. The *X*- and *Y*-axes represent the vertical and horizontal axes from Figure 1. The *Z*-axis represents the zoom level, thereby illustrating where on the canvas the users were interacting and at what depth those interactions were occurring. As the plots show, in the tunnel condition the participants exhibited similar patterns of behavior, but that behavior was concentrated around different points on the canvas.

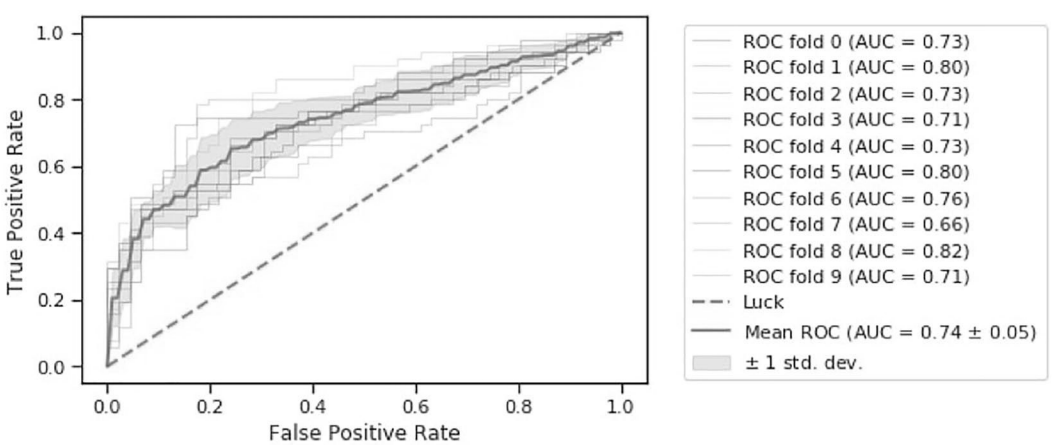


Figure 4. ROC curves for 10-fold cross validation of the CNN-LSTM model. AUC = area under curve; ROC = relative operating characteristics; CNN = Convolutional Neural Network; LSTM = long short-term memory.

feature-sparse dataset. This shows promise for adaptive systems based on PDM.

Direct comparison of classifier performance is difficult as there are relatively few benchmarks and the degree to which an attentional state manifests behaviorally is likely contingent on the characteristics of the attentional state itself as well as the task that users are performing while in that state. Our main point of comparison was the Régis et al. (2014) study, as they successfully demonstrated classification of attentional tunnels through largely physiological measures. They achieved higher classification accuracy than we did (91%). However, the

differences between experimental conditions were much more overt in their study, so direct comparison is difficult. Mannaru, Balasingam, Pattipati, Sibley, and Coyne (2016) used physiological measures (eye tracking and pupillary dilation) to measure workload in operators using a desktop task. Again, they achieved higher classification accuracy (88%) but with overt differences between conditions. McDonald et al. (2014) represents an interesting comparison as they also used behavioral metrics, but to infer states of drowsiness in drivers. We achieved a higher classification accuracy than they did as measured by AUC (AUC = 0.70), but because their attentional state and task were very different, it is difficult to compare. These comparisons demonstrate that PDM can yield results that are significantly above chance and comparable with physiological measurement and should therefore be considered when implementing adaptive user interfaces triggered by operator state measurement.

Although the dataset that we trained our classifier on was sufficient to achieve comparable performance, it still had a high degree of variability from participant-to-participant, was relatively small, and included few features. Future studies should examine the performance of similar classifiers trained over longer time spans on a more homogeneous participant pool. Such conditions might better approximate those of actual work settings and operator characteristics. Furthermore, as we wanted to evaluate the efficacy of PDM, we only passed interaction data to the classifier. If we were to include system state data or further contextual cues, the accuracy of the classifier would presumably improve. Including these additional features is feasible in many information dense spaces as system state information is often readily available and time-stamped (e.g., cybersecurity; Corchado & Herrero, 2011). Finally, we used relatively short time windows (18.75 s). With a sufficiently large dataset, larger time windows could be used, which would also likely increase the accuracy of the classifier.

GENERAL DISCUSSION

This study supports the efficacy of using PDM for inferring attentional states. We were able to validate that the experimental paradigm successfully elicited behaviors that aligned with the Wickens and Alexander (2009) definition of an attentional tunnel while also demonstrating the value of refocusing that definition on the *potential for* rather than the *expectation of* negative outcomes. We then developed a classifier capable of inferring attentional tunnels from data remotely recorded through Amazon's Mechanical Turk. Given the unknown locations of the users, the computers they used to complete the study, and the demographics of the users themselves, this study demonstrates the potential and practicality of PDM in adaptive automation.

The performance trade-off that we found shows that fixating on a particular area of space should not necessarily be avoided, which indicates that adaptive DSSs should not be designed to "break" attentional tunnels. Rather, depending on the domain, they should be designed to foster an understanding of the stakeholders' attentional processes. For example, a shift supervisor may want to know when their operators are in attentional tunnels so that they can determine the appropriateness of that state. Future studies can also seek to develop automated methods of distinguishing between attentional tunnels and directed attention. Alternatively, the methods described in this paper can be used to determine how susceptible an interface is to inducing attentional tunnels.

To further improve the utility of this classifier in practice, system state information can be included as either additional features or classifier modulators. For example, Rantanen and Goldberg (1999) demonstrated that increased mental workload narrows an operator's visual field size. If workload was coupled with this classifier, the classification threshold could be modified to facilitate more accurate classifications. Future studies can also examine this research through the lens of the Strategic Task Overload Model (STOM; Wickens, Gutzwiller, & Santamaria, 2015), which may be able to formalize how and when people switch between areas of interest during visual search. This is particularly relevant as attentional tunneling has been cited as a critical aspect of STOM (Wickens & Gutzwiller, 2017).

CONCLUSION

Practical applications of adaptive automation have remained sparse due to an inability to accurately capture the context of the adaptations (Feigh et al., 2012). However, recent developments in machine learning may offer a solution to this problem by enabling a richer contextual understanding (Mangos & Hulse, 2017). We show here that combining some of these new methods with PDM can allow for one dimension of attentional context to be understood. Future work should examine the suitability of this approach in combination with different aspects of operational context and for a wide array of attentional states.

For access to either our dataset or the source code for a CogLog application, please contact the corresponding author Sean W. Kortschot at sean.kortschot@mail.utoronto.ca.

ACKNOWLEDGMENTS

This project was partially supported by Uncharted Software Inc., the Ontario Centers of Excellence (Grant No. 23884), and the Natural Sciences and Engineering Research Council (Grant No. CRDPJ 500832-16). We thank Scott Langevin and Scott Ralph at Uncharted Software as well as Amrit Prasad at The University of Toronto for their contributions.

KEY POINTS

- There is a performance trade-off stemming from attentional tunnels: Improved performance when information appears within a tunnel and impaired performance when information appears outside of that tunnel.
- Attentional tunnels manifest in predictable patterns of behavior. These patterns can be detected passively and classified using machine learning methods.
- Passive Data Monitoring is a promising alternative to physiological measurement of operator attentional state.
- The appropriateness of an attentional state is dynamic. As such, successful adaptive automation will require greater consideration of the attentional demands of users as well as the state of the system.

REFERENCES

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A public domain dataset for human activity recognition using smartphones. In *21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges, Belgium. Retrieved from https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2013-84.pdf

Arif, M., & Kattan, A. (2015). Physical activities monitoring using wearable acceleration sensors attached to the body. *PLOS ONE*, *10*, 1–16. doi:10.1371/journal.pone.0130851

Atchley, P., & Dressel, J. (2004). Conversation limits the functional field of view. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*, 664–673. doi:10.1518/hfes.46.4.664.56808

Baldwin, C. L., Roberts, D. M., Barragan, D., Lee, J. D., Lerner, N., & Higgins, J. S. (2017). Detecting and quantifying mind wandering during simulated driving. *Frontiers in Human Neuroscience*, *11*, 1–15. doi:10.3389/fnhum.2017.00406

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., & Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, *78*, B231–B244. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17547324

Borghetti, B. J., Giametta, J. J., & Rusnock, C. F. (2017). Assessing continuous operator workload with a hybrid scaffolded neuroergonomic modeling approach. *Human Factors*, *59*, 134–146. doi:10.1177/0018720816672308

Briggs, G. F., Hole, G. J., & Turner, J. A. J. (2018). The impact of attentional set and situation awareness on dual tasking driving performance. *Transportation Research Part F: Traffic Psychology and Behaviour*, *57*, 36–47. doi:10.1016/j.trf.2017.08.007

Budd, A., & Björklund, E. (2016). Cutting-edge visual effects. In A. Budd (Ed.), *CSS mastery* (pp. 335–370), Berkeley, CA: Apress.

Corchado, E., & Herrero, Á. (2011). Neural visualization of network traffic data for intrusion detection. *Applied Soft Computing Journal*, *11*, 2042–2056. doi:10.1016/j.asoc.2010.07.002

Dixon, B. J., Daly, M. J., Chan, H., Vescan, A. D., Witterick, I. J., & Irish, J. C. (2013). Surgeons blinded by enhanced navigation: The effect of augmented reality on attention. *Surgical Endoscopy*, *24*, 454–461. doi:10.1007/s00464-012-2457-3

Dongen, K., & Van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *Journal of Human Computer Studies*, *71*, 410–424. doi:10.1016/j.ijhcs.2012.10.018

Dorneich, M. C., Whitlow, S. D., Mathan, S., Carciofini, J., & Ververs, P. M. (2005). The communications scheduler: A task scheduling mitigation for a closed loop adaptive system. In D. D. Schmorrow (Ed.), *Foundations of augmented cognition* (pp. 132–141).

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*, 1008–1024. doi:10.1177/0018720812443983

Friedman, R. S., Fishbach, A., Förster, J., & Werth, L. (2003). Attentional priming effects on creativity. *Creativity Research Journal*, *15*, 277–286. doi:10.1080/10400419.2003.9651420

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *Transactions on Neural Networks and Learning Systems*, *28*, 2222–2232. Retrieved from http://arxiv.org/abs/1503.04069

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2018). Multivariate LSTM-FCNs for time series classification. *IEEE Access*. Retrieved from http://arxiv.org/abs/*1801*.04503

Kortschot, S. W., Sovilj, D., Jamieson, G. A., Sanner, S., Carrasco, C., & Soh, H. (2018). Measuring and mitigating the costs of attentional switches in active network monitoring for cybersecurity. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*, 962–977. http://doi.org/10.1177/0018720818784107

Kortschot, S. W., Sovilj, D., Soh, H., Jamieson, G. A., Sanner, S., Carrasco, C., & Ralph, S. (2017). An open source adaptive user interface for network monitoring. In *2017 IEEE International Conference on Systems, Man, and Cybernetics*. Banff, Alberta, Canada. Retrieved from https://ieeexplore.ieee.org/document/8122832

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. L. Damos (Ed.), *Multiple task performance* (pp. 279–328). Bristol, PA: Taylor & Francis.

Lee, Y. Y., & Hsieh, S. (2014). Classifying different emotional states by means of EEG-based functional connectivity patterns. *PLOS ONE*, *9*(4), e95415. doi:10.1371/journal.pone.0095415

Li, W., Mo, W., Zhang, X., Squiers, J. J., Lu, Y., Sellke, E. W., & Thatcher, J. E. (2015). Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging. *Journal of Biomedical Optics*, *20*, 1–9. doi:10.1117/1.JBO.20.12.121305

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). *A critical review of recurrent neural networks for sequence learning*. Retrieved from https://arxiv.org/abs/*1506*.00019

Mac Aoidh, E., Bertolotto, M., & Wilson, D. C. (2012). Towards dynamic behavior-based profiling for reducing spatial information overload in map browsing activity. *Geoinformatica*, *16*, 409–434. doi:10.1007/s10707-011-0137-4

Mangos, P. M., & Hulse, N. A. (2017). Advances in machine learning applications for scenario intelligence: Deep learning. *Theoretical Issues in Ergonomics Science*, *18*, 184–198. doi:10.1080/1463922X.2016.1166406

Mannaru, P., Balasingam, B., Pattipati, K., Sibley, C., & Coyne, J. (2016). Cognitive context detection for adaptive automation. *Proceedings of the Human Factors and Ergonomics Society*, *60*, 223–227. doi:10.1177/1541931213601050

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*, 2145–2166. doi:10.1256/003590002320603584

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2014). Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures. *Human Factors*, *56*, 986–998. doi:10.1177/0018720813515272

McLane, H. C., Berkowitz, A. L., Patenaude, B. N., Mckenzie, E. D., Wolper, E., Wahlster, S., & Mateen, F. J. (2015). Availability, accessibility, and affordability of neurodiagnostic tests in 37 countries. *Neurology*, *85*, 1614–1622. doi:10.1212/WNL.0000000000002090

Mills, K. C., Spruill, S. E., Kanne, R. W., Parkman, K. M., & Zhang, Y. (2001). The influence of stimulants, sedatives, and fatigue on tunnel vision: Risk factors for driving and piloting. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*, 310–327. http://doi.org/10.1518/001872001775900878

Mukhopadhyay, S. C., & Lay-Ekuakille, A. (Eds.). (2010). *Advances in biomedical sensing, measurements, instrumentation, and systems*. Berlin, Germany: Springer.

Pak, U., Kim, C., Ryu, U., Sok, K., & Pak, S. (2018). A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. *Air Quality, Atmosphere & Health*, *11*, 883–895.

Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., & De Vos, M. (2016). Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, *64*, 1761–1771.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. doi:10.1007/s13398-014-0173-7.2

Prieto, O. J., Alonso-González, C. J., & Rodríguez, J. J. (2015). Stacking for multivariate time series classification. *Pattern Analysis & Applications*, *18*, 297–312. doi:10.1007/s10044-013-0351-9

Rantanen, E. M., & Goldberg, J. H. (1999). The effect of mental workload on the visual field size and shape. *Ergonomics*, *42*, 816–834. doi:10.1080/001401399185315

Régis, N., Dehais, F., Rachelson, E., Thooris, C., Pizziol, S., Causse, M., & Tessier, C. (2014). Formal detection of attentional tunneling in human operator-automation interactions. *IEEE Transactions on Human-machine Systems*, *44*, 326–336. doi:10.1109/THMS.2014.2307258

Rogé, J., Kielbasa, L., & Muzet, A. (2002). Deformation of the useful visual field with state of vigilance, task priority, and central task complexity. *Perceptual and Motor Skills*, *95*, 118–130. doi:10.2466/pms.2002.95.1.118

Rubenstein, E., & Mason, J. F. (1979). An analysis of Three Mile Island: The accident that shouldn't have happened. *IEEE Spectrum*, *13*, 33–42.

Shappell, S. A, & Wiegmann, D. A. (2003). *A human error analysis of general aviation controlled flight into terrain accidents occurring between 1990-1998* (Final report). Washington, DC. Retrieved from https://trid.trb.org/view/659810

Tan, J. H., Hagiwara, Y., Pang, W., Lim, I., Oh, S. L., Adam, M., & Acharya, U. R. (2018). Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Computers in Biology and Medicine*, *94*, 19–26. doi:10.1016/j.compbiomed.2017.12.023

Verwey, W. B., Shea, C. H., & Wright, D. L. (2015). A cognitive framework for explaining serial processing and sequence execution strategies. *Psychonomic Bulletin & Review*, *22*, 54–77. doi:10.3758/s13423-014-0773-4

Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2015). A review of human activity recognition methods. *Frontiers in Robotics and AI*, *2*, 28. doi:10.3389/frobt.2015.00028

Wickens, C. D., & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *The International Journal of Aviation Psychology*, *19*, 182–199. doi:10.1080/10508410902766549

Wickens, C. D., & Gutzwiller, R. S. (2017). The status of the strategic task overload model (STOM) for predicting multi-task management. In *Proceedings of the Human Factors and Ergonomics Society*, *61*, 757–761. doi:10.1177/1541931213601674

Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A meta-analyses and a model.

Journal of Human Computer Studies, *79*, 79–84. doi:10.1016/j.ijhcs.2015.01.002

Williams, L. J. (1995). Visual field narrowing induced by workload. *The Journal of General Psychology*, *122*, 225–235.

Wilson, G. F., & Russell, C. A. (2004). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*, 381–389. doi:10.1518/hfes.45.3.381.27252

Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, *11*, 68–75. doi:10.1049/iet-its.2016.0208

Sean W. Kortschot is a PhD candidate at the University of Toronto. He received a master's degree in applied science from the University of Toronto in 2016.

Greg A. Jamieson is a professor and Clarice Chalmers Chair of engineering design at the University of Toronto. He received his PhD in mechanical and industrial engineering from the University of Toronto in 2003. He is a licensed professional engineer in Ontario.

Date received: October 25, 2018
Date accepted: May 17, 2019

Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming

Na Du, **Kevin Y. Huang**, and **X. Jessie Yang**, University of Michigan, Ann Arbor, USA

Objective: The study examines the effects of disclosing different types of likelihood information on human operators' trust in automation, their compliance and reliance behaviors, and the human-automation team performance.

Background: To facilitate appropriate trust in and dependence on automation, explicitly conveying the likelihood of automation success has been proposed as one solution. Empirical studies have been conducted to investigate the potential benefits of disclosing likelihood information in the form of automation reliability, (un)certainty, and confidence. Yet, results from these studies are rather mixed.

Method: We conducted a human-in-the-loop experiment with 60 participants using a simulated surveillance task. Each participant performed a compensatory tracking task and a threat detection task with the help of an imperfect automated threat detector. Three types of likelihood information were presented: overall likelihood information, predictive values, and hit and correct rejection rates. Participants' trust in automation, compliance and reliance behaviors, and task performance were measured.

Results: Human operators informed of the predictive values or the overall likelihood value, rather than the hit and correct rejection rates, relied on the decision aid more appropriately and obtained higher task scores.

Conclusion: Not all likelihood information is equal in aiding human-automation team performance. Directly presenting the hit and correct rejection rates of an automated decision aid should be avoided.

Application: The findings can be applied to the design of automated decision aids.

Keywords: human–robot interaction, trust in automation, likelihood alerts, Bayesian inference, base rate fallacy

Address correspondence to X. Jessie Yang, Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109, USA; e-mail: xijyang@umich.edu.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 987–1001
DOI: 10.1177/0018720819862916
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Automated decision aids have been used in a wide array of domains, including military operations, medical diagnosis, transportation safety administration (TSA), among others. As automation becomes more capable in perception, planning, learning, and action execution, it is expected to significantly enhance the human-automation team performance. However, issues arise when human agents place unjustified trust in and dependence on automation or when they do not display enough trust and dependence (Dixon, Wickens, & McCarley, 2007; Du et al., 2019; Lee & See, 2004; Parasuraman & Riley, 1997; Petersen, Robert, Yang, & Tilbury, 2019; Yang, Unhelkar, Li, & Shah, 2017).

To facilitate appropriate trust in and dependence on automation, explicitly conveying the likelihood of automation success has been proposed as one solution. Empirical studies have investigated the potential benefits of disclosing likelihood information in the form of automation reliability, (un)certainty, and confidence. Among existing studies, few were based on specific computational algorithms, for instance, the neural network used in a study by McGuirl and Sarter (2006). Not surprisingly, to model the performance of the automation, the majority of existing studies used the signal detection theory (SDT; Macmillan & Creelman, 2005; Tanner & Swets, 1954), based on which the likelihood information is calculated. Yet, results from these studies seem to be inconsistent. Some studies revealed that the likelihood information significantly helped human operators calibrate their trust, adjust their reliance and compliance behaviors, and enhance human-automation team performance (McGuirl & Sarter, 2006; Walliser,

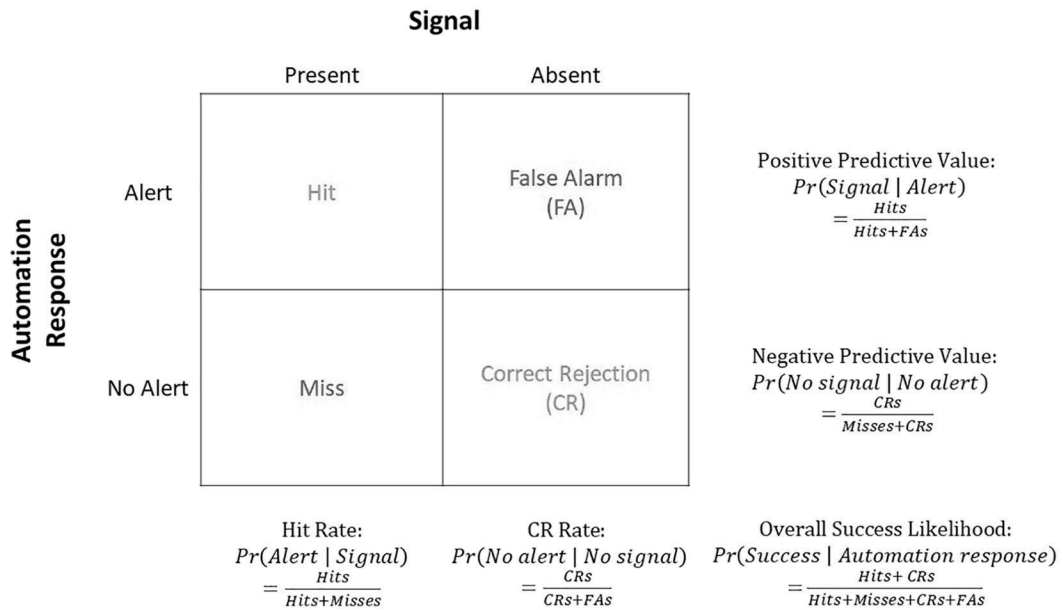


Figure 1. Signal detection theory and calculations of hit rate, correct rejection rate, positive predictive value, negative predictive value, and overall success likelihood.

de Visser, & Shaw, 2016; Wang, Jamieson, & Hollands, 2009). Other studies, however, reported that human operators did not trust or depend on automated decision aids appropriately even when the likelihood information was disclosed (Bagheri & Jamieson, 2004; Fletcher, Bartlett, Cockshell, & McCarley, 2017). A close examination of existing literature suggests that studies use different methods to calculate the likelihood information, which potentially contribute to the mixed results.

SDT models the relationship between signals and noise, as well as the automation's ability to detect signals among noise. The state of the world is characterized by either "signal present" or "signal absent," which may or may not be identified correctly by the automation. The combination of the state of the world and the automation's detection results in four possible states: hit, miss, false alarm (FA), and correct rejection (CR; see Figure 1).

Based on the framework of SDT, the calculation of automation likelihood information can be broadly classified into three categories. The first category of likelihood information is the automation's overall likelihood of success regardless of hits or CRs, calculated as Pr(Success|Automation Response)=(Hits + CRs)/(Hits + Misses + FAs + CRs). For example, Dzindolet, Pierce, Beck, and Dawe (2002) examined how revealing the number of errors an automated decision aid made affected the perceived performance of and reliance on the automated aid. In their study, participants viewed 200 slides of displaying pictures of a military terrain and indicated whether or not a soldier in camouflage was in the slide with the help from either an automated decision aid or a human decision aid. After 200 trials, half of the participants were provided with the reliability of the decision aid (total number of errors) and the other half not. The participants then rated the decision aid's performance and indicated whether to rely on the aid for the target detection task in 10 trials randomly chosen from the past 200 trials. Results showed that both types of decision aids were rated more favorably when its reliability was disclosed. More recently, Walliser et al. (2016) conducted a study where participants interacted with four unmanned aerial vehicles (UAVs) that used automated target recognition (ATR) systems to identify targets as enemy or friendly. Results showed that when participants were informed of the overall success

likelihood information ("corrected identification rate" in the article), participants tended to apply a more appropriate strategy when interacting with the automation, resulting in better task performance.

The second type of likelihood information is the predictive value, calculated as Hits/(Hits/+FAs) or CRs/(Misses + CRs). The positive predictive value means the probability of having a true signal given an automation alert, Pr(Signal|Alert), and the negative predictive value means the probability of not having a signal when the automation is silent, Pr(No signal|No alert). Along this line of research, Wang et al. (2009) examined the effects of presenting the positive predictive value on human operators' belief, trust, and dependence using a combat identification (CID) task. In the study, participants distinguished friend from foe with the aid from an imperfect CID. More specifically, due to its working mechanism, once the CID identified a soldier as friendly, it was always correct. However, when the CID identified a soldier as "unknown," the solider could be "friendly," "hostile," or "neutral." Half of the participants were informed of the positive predictive value and the other half not. Results of their study revealed that disclosing the positive predictive value to users positively influenced trust and reliance. In a follow-up study, Neyedli, Hollands, and Jamieson (2011) developed four visual displays for presenting predictive values in the CID task. Display type (pie, random mesh) and display proximity (integrated, separated) of likelihood information were manipulated in the experiment. The results revealed that participants relied on the automation more appropriately and had greater sensitivity with the integrated display and random mesh display. Studies on likelihood alarms also shed light on the effects of disclosing the predictive values. Likelihood alarms, in contrary to traditional binary alarms, integrate both state information and likelihood information by dividing a state into two or more graded levels. For instance, "warning" and "caution" both could indicate the presence of a target, with "warning" indicating a higher probability. Although not explicitly stated, these studies manipulated the positive and negative predictive values to represent the varying likelihood of true positives and true negatives given the automation responses, showing that in general, human operators demonstrated higher trust in and dependence on alerts with higher likelihood (Sorkin, Kantowitz, & Kantowitz, 1988; Wiczorek & Manzey, 2014; Yang et al., 2017). Despite the above-mentioned positive evidence, Fletcher et al. (2017) asked participants to view a series of simulated sonar returns and decide whether a target was present or not. However, the display of rings indicating the likelihood that a target was present, given a return signal, did not seem to improve the overall ability of participants to distinguish targets from noise.

The third type of likelihood information is the hit rate and CR rate, calculated as Hits/(Hits + Misses) and CRs/(FAs + CRs). Hit rate is the probability of automation issuing an alarm or an alert given a true signal, Pr(Alert|Signal), and CR rate is the probability of automation silence when there is no signal, Pr(No alert|No Signal). It is important to distinguish the predictive values from the hit/CR rates. In fact, the positive/negative predictive values and the hit/CR rates are *inverse* conditional probabilities of each other. The two predictive values can be derived from the hit/CR rates using the Bayes theorem (see the "The Present Study" section for details). Utilizing the Multi-Attribute Task Battery (MAT; Comstock & Arnegard, 1992), Bagheri and Jamieson (2004) examined the effect of providing operators with information about the context-related nature of automation reliability. Participants performed three tasks simultaneously: tracking, fuel management, and system monitoring. The monitoring task was automated and a gauge showing abnormal numbers would automatically reset its value. However, sometimes the automation would fail (miss) to correct the value and the human operator should intervene. Automation reliability, essentially hit rate ("slightly below 100%" for high hit rate or "slightly above 50%" for low hit rate), was disclosed to the participants. Comparing with a previous study where participants were unaware of the likelihood information, there seemed to be no evidence on any beneficial effects of disclosing hit rate on trust in automation or task performance.

THE PRESENT STUDY

The above-mentioned studies on likelihood information suggest that disclosing the overall

likelihood information could increase preference and task performance. In addition, in general, there is positive evidence supporting that presenting predictive values could help human operators calibrate their trust and adjust their dependence behaviors, leading to better performance. In contrast, revealing hit/CR rates does not seem to be beneficial. Despite the inconsistent results, there is little, if not no, research directly comparing the effects of revealing different types of likelihood information.

In the present study, we aimed to investigate whether and how different methods of calculating likelihood information affect operators' trust in and dependence on automation, and task performance. We argue that the beneficial effects of disclosing the likelihood information are influenced by, at least, two factors. The first factor is information granularity—the extent to which the likelihood information represents probabilistic information specific to certain conditions. The overall success likelihood, Pr(Success|Automation response), is less fine-grained compared with the predictive values and the hit/CR rates, as it represents an aggregated probability, regardless what the automation response is (alert or no alert). The second factor is information directness—the extent to which the likelihood information can be directly used to guide people's behaviors without the need to estimate or integrate other information. The predictive values are the most direct in guiding people's compliance and reliance behaviors. The positive predictive value, Pr(Signal|Alert = x%), indicates that when the automation's alert or alarm goes off, there is x% chance that there is a true signal. Probabilistically speaking, if the automation's alarm goes off 100 times, there would be x true alarms and $100 - x$ FAs. And an optimal decision maker should only check the x number of true alarms and save his or her time and resources when FAs happen. The same logic applies to the negative predictive value. On the contrary, the hit/CR rates, Pr(Alert|Signal) and Pr(No alert|No Signal), are less usable, because the human operator cannot directly use the probabilities to guide their behaviors. Instead, the hit/CR rates need to be integrated with the base rate to generate useful information in guiding behaviors. And this particular integration process, known as Bayesian inference, is very difficult (Kahneman, 2011). To better illustrate the idea of Bayesian inference, consider the following scenario:

> An airport security officer detects threats with the help of an nearly perfect decision aid. The alarm of the decision aid goes off if it recognizes a threat. The security officer could also manually check any luggage. The decision aid is correct 95% of the time. In other words, if there is a threat, the decision aid recognizes it with a 95% probability (Hit rate = Pr(Alarm|Threat) = 95%), and if there is no threat, the aid shows no threat with a 95% probability (CR rate = Pr(No alarm|No threat) = 95%). Suppose threats are rare in the airport, on average occurring only 1% of the time. If an alarm went off, should the officer panic and what would be the chance that there was actually a threat?

In the example, the hit rate is 95%. However, it does not mean that when an alarm goes off, there is 95% chance that there would be a threat. To answer the question correctly, we need to apply the Bayes's rule to calculate the positive predictive value, mathematically the *inverse* of the hit rate:

$$\Pr(Threat|Alarm) = \frac{\Pr(Alarm|Threat)\Pr(Threat)}{\Pr(Alarm)}$$

$$= \frac{\Pr(Alarm|Threat)\Pr(Threat)}{\Pr(Alarm|Threat)\Pr(Threat) + \Pr(Alarm|No\ threat)\Pr(No\ threat)}$$

$$= \frac{95\% \times 1\%}{95\% \times 1\% + 5\% \times 99\%} = 16\%.$$

The probability of a true threat is only 16%! If we do not consider the payoff structure associated with the task (i.e., high cost if missing a threat), the result indicates that probabilistically the officer only needs to manually check 16 luggage out of 100 alarms and could invest his or her time on other tasks 84% of the time when alarms go off.

Prior research shows that it is cognitively demanding to use the Bayes's rule (Bar-Hillel, 1980; Cosmides & Tooby, 1996; Goodie &

Fantino, 1996; Kahneman, 2011) because of several reasons. First, the base rate may not be readily available and an operator needs to estimate it. Second, when making a probabilistic judgment, an operator may neglect the base rate of Pr(Threat), that threats only occur 1% of the time (Kahneman, 2011). Third, a person might be confused about Pr(Alram|Threat) and its *inverse*, Pr(Threat|Alarm), as both are related to the probability of an accurate threat identification (Bar-Hillel, 1980). Due to the difficulty in performing Bayesian inference, we speculate that the hit/CR rates are the least direct.

The overall success likelihood, Pr(Success|Automation respones), represents the probability of a true state (hit or CR), given an automation response (alert or no alert), and a higher probability means an operator should follow the automation more overall. The overall success likelihood alone only guides human operators' behaviors at an aggregated level—if the automation overall success likelihood is 80%, when the automation issues 100 suggestions (regardless what the suggestion is), 80 suggestions are correct. Despite the lack of granularity, we speculate that the overall likelihood information is more direct than the hit/CR rate, as it can be easily used to guide overall human behaviors.

Due to the influence of the two factors, we predicted that there would be significant differences in participants' trust, dependence, and dual-task performance when presented with different types of likelihood information. In particular, disclosing hit/CR rate would be the least beneficial in fostering proper trust and dependence, and would lead to the worst task performance. Revealing the predictive values, in contrast, would be the most beneficial.

METHOD

This research complied with the American Psychological Association code of ethics and was approved by the institutional review board at the University of Michigan.

Participants

A total of 25 male and 36 female university students (average age = 22.28 years, *SD* = 4.88) with normal or corrected-to-normal vision participated in the experiment. Participants were compensated with US$10 upon completion of the experiment. In addition, there was a chance to obtain an additional bonus of 1 to 5 dollars based on their performance.

Apparatus and Stimuli

We used a simulated surveillance task in the experiment. In the experimental task, participants were asked to control the level of flight of a simulated swarm of drones, essentially a compensatory tracking task, and simultaneously detect potential threats in photo feeds from the drones (Figure 2). Participants were only able to access the display for either the tracking task or the detection task at any time and needed to toggle between the two displays. The simulated surveillance task was programmed using Java and the experiment was run on a 24 in. monitor.

Tracking task. Each trial started on the tracking display and lasted 10 s. The tracking task was programmed based on the PEBL (The Psychology Experiment Building Language) compensatory tracker task (http://pebl.sourceforge.net/battery.html). Participants used a joystick to move a randomly drifting green circle to a crosshair located at the center of the screen—that is, minimize the distance between the green circle and the crosshair as shown in Figure 2(a). When a trial started, the green circle started at the center of the crosshair. The position of the circle is a function of its previous position, its velocity, and the actions of three forces. The first force is the user input. The second force is a buffeting force composed of six sine waves at different amplitudes, frequencies, and phase angles. The third force simulates the force of gravity that causes the circle to slip on an unseen slippery surface. As a result of the buffeting force and the gravitational force, the circle drifts randomly. The performance of the tracking task is measured by two metrics: the root mean square of the tracking errors (RMSE) and the tracking score ranging 0 to 10. The tracking error—the distance in pixels between the location of the circle and the crosshair—was measured at a frequency of 20 Hz. The RMSE was calculated as $\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(\text{tracking error})^2}$, where $n = 200$. The tracking score was calculated using a 10-bin histogram of the RMSE distribution based on a

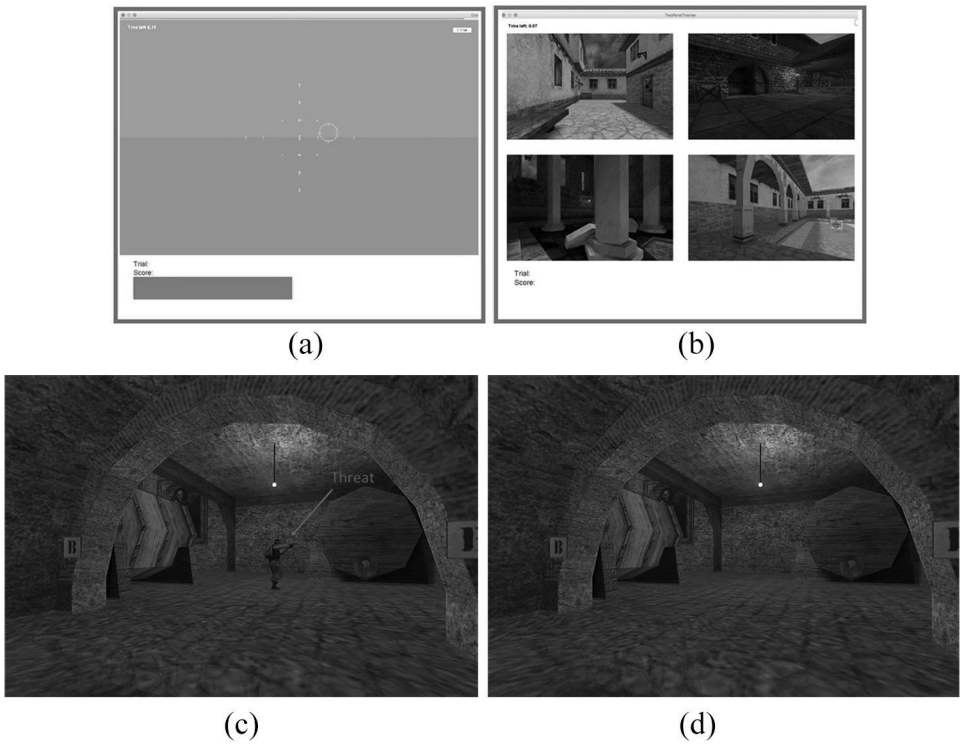


Figure 2. Dual-task environment in the simulation testbed: (a) tracking task, (b) detection task, (c) threats.

dataset collected in a prior study (Yang et al., 2017).

Detection task. Besides the tracking task, every trial participants received a new set of four images from the simulated drones and were responsible for threat detection. The four images were static during every trial as shown in Figure 2(b). The threat was a person as shown in Figure 2(c) and only one threat would present in one of the four images. There was no distractor in the four images and the participants did not determine whether the person was a friend or a foe. The distribution of the threats across the four images followed a uniform distribution. Participants performed the detection task with the help of an imperfect automated threat detector. If the detector recognized a threat, the alert "Danger" went off immediately when a trial started in both visual and auditory modalities. The visual red alert was only shown on the tracking display (Figure 2(a)) and the auditory notification was a synthetic sound of "Danger." Participants were expected to identify the presence of the threat by pressing the "Report" button on the joystick as accurately and quickly as possible. The participant could follow the decisions of the threat detector blindly, or check the images in person and make his or her own decisions. If the detector identified no threat, the alert was silent. Participants did not report the absence of threat, that is, participants were expected to perform no action when there was no threat. The performance of the detection task is measured by detection time, detection accuracy, and detection score (see the "Scoring System" section).

Toggle between two displays. Every trial started on the tracking display. Participants were only able to access one display at a time, and needed to toggle between the displays of the tracking and the detection tasks using a "Switch" button on the joystick. There was a 0.5-s time delay every time they toggled between the displays, simulating the time for computer processing and loading the displays. The time stamp and the number of occurrences of participants

TABLE 1: Corresponding Hits, Misses, False Alarms, and Correct Rejections

Reliability	*c*	*d′*	Alert	Threat	No Threat
Low	–0.25	1.5	Danger	13	11
			Clear	2	24
High	–0.25	3	Danger	14	4
			Clear	1	31

pressing the "Switch" button were tracked automatically by the program.

Scoring system. In the experimental task, participants performed the tracking task and the detection task simultaneously, and needed to make a trade-off decision on which task to perform at any time, that is, if they decide to check the four images, they would probably earn more points in the detection task but fewer points in the tracking task, and vice versa. Therefore, a payoff structure has to be determined to eliminate potential bias toward either the tracking or the detection task by ensuring that the potential gain in one task is approximately equal to the opportunity cost in the other task. A pilot study was conducted to determine the payoff structure (see the appendix for more details). As a result, every trial participants could obtain 0 to 10 points for the tracking task and 0 to 5 points for the detection task:

$$\text{Detection score} = \begin{cases} 0 & \text{Detection is wrong} \\ 5 - 5 \times \left(\dfrac{\text{detection time}}{10{,}000\ \text{ms}} \right) & \text{Detection is correct: Hit} \\ 5 & \text{Detection is correct: CR} \end{cases}.$$

Experimental Design

The experiment adopted a 2 (automation reliability: low vs. high) × 3 (likelihood information: overall success likelihood, predictive values, and hit/CR rates) mixed design with automation reliability as the within-subjects factor and likelihood information as the between-subjects factor.

The reliability of the automated threat detector was configured based on SDT. In the present study, the criterion *c* was set at –0.25 and sensitivity *d′* at 1.5 or 3, resulting in automation with low and high reliability (Table 1). Benchmarking prior literature (McBride, Rogers, & Fisk, 2011; Wiczorek & Manzey, 2014; Yang et al., 2017), we set the base event rate at 30%. Based on the preset *c, d′*, and base rate, the number of occurrences of hits, misses, FAs, and CRs were calculated and rounded to integers.

Different types of likelihood information were calculated as follows:

$$\begin{aligned} \text{Overall success likelihood} &= \frac{\text{Hits}+\text{CRs}}{\text{Hits}+\text{Misses}+\text{FAs}+\text{CRs}} \\ &= 74\% \text{ or } 90\%, \\ \text{Positive predictive value} &= \frac{\text{Hits}}{\text{Hits}+\text{FAs}} = 54\% \text{ or } 78\%, \\ \text{Negative predictive value} &= \frac{\text{CRs}}{\text{Misses}+\text{CRs}} = 92\% \text{ or } 97\%, \\ \text{Hit rate} &= \frac{\text{Hits}}{\text{Hits}+\text{Misses}} = 87\% \text{ or } 93\%, \\ \text{Correct rejection rate} &= \frac{\text{CRs}}{\text{FAs}+\text{CRs}} = 69\% \text{ or } 89\%. \end{aligned}$$

Measures

Trust. We measured participants' subjective rating of trust using a visual analog scale (Wiczorek & Manzey, 2014). The leftmost anchor of the trust scale indicated, "I don't trust the detector at all" and the rightmost anchor, "I absolutely trust the detector." The visual analog scale was then converted to a 0 to 100 scale. As part of the testbed design, in addition to trust ratings, participants needed to report their self-confidence in performing the task without the decision aid and perceived reliability of the decision aid. As the two measures were less relevant to this study, we did not report the data analysis results.

Compliance and reliance. We also assessed participants' compliance and reliance behaviors. Compliance and reliance were operationally defined as the possibility of a participant blindly following the recommendation given by

the automated threat detector without cross-checking the detection display. In particular, compliance was calculated as the possibility that a participant blindly reports a threat upon receiving a "Danger" alert without cross-checking the detection display, and reliance was calculated as the possibility that the participant neither reports nor cross-checks when the detector is silent:

$$\text{Compliance} = \Pr\left(\begin{array}{l}\text{Report AND not cross-}\\ \text{checking}|\text{Alert}\end{array}\right),$$

$$\text{Reliance} = \Pr\left(\begin{array}{l}\text{Not reporting AND not}\\ \text{cross-checking}|\text{ No alert}\end{array}\right).$$

Performance. The performance of the detection task was measured by the detection accuracy and detection time, as well as the detection score. The performance of the tracking task was calculated using the RMSE and the tracking score. The combined performance of both tasks was calculated as the sum of the detection score and the tracking score.

Experimental Procedure

Upon arrival, participants provided informed consent and filled out a demographics form. Afterward, participants received practice on the tasks. The practice session consisted of a 30-trial block with the tracking task only and an eight-trial block of combined tasks, where participants experienced two hits, two misses, two FAs, and two CRs. Participants were informed that the automated threat detector used in the practice was just for illustration purpose. Afterward, they were randomly assigned to one of the three likelihood information conditions. A table similar to Figure 1 was then shown to the participants. Based on the condition a participant was assigned to, the definition, the meaning, and the calculation of a particular likelihood information were introduced to the participant. To ensure that participants understood the likelihood information, the participants were given an example with different number of hits, misses, FAs, and CRs, and were asked to calculate the likelihood information themselves. If a participant had difficulty doing so, the verbal definitions were reiterated and shown again to the participant, with potential further clarification on specific terms, until the correct answer was reached by the participant.

The experiment consisted of two 50-trial blocks with different automation reliability. The order of automation reliability was counterbalanced. Participants were verbally informed of the values of the likelihood information prior to the experiment. A text message showing the probability was also present throughout the experiment. Prior to the onset of each trial, there was a splash screen with a 3-s countdown timer. After every trial, participants were informed of the detection accuracy, the tracking score, and the detection score they obtained in this trial and the accumulative scores they had obtained so far. After every five trials, participants indicated their trust. Participants were told that their subjective ratings should be based on all the trials they have completed so far, instead of just the previous five trials.

RESULTS

Data from one participant were excluded from analysis as his tracking task performance was below three standard deviations from the mean. All hypotheses were tested using data from the remaining 60 participants. We used mixed-design analysis of covariance (ANCOVA) to analyze the relationship between independent variables and dependent variables. Participants' tracking task performance (last 10 trials) in the practice session was used as the covariate for analysis. The α level was set at .05 for all statistical tests. All post hoc comparisons used a Bonferroni α correction.

Subjective Trust

Trust. Participants had higher trust in the automated threat detector as automation reliability increased, $F(1, 56) = 7.533$, $p = .008$. However, the effect of likelihood information was nonsignificant.

Compliance and Reliance

Figure 3 shows the participants' compliance and reliance behaviors.

Compliance. Higher automation reliability led to higher compliance rate on the automated threat detector, $F(1, 56) = 7.196$, $p = .01$. The

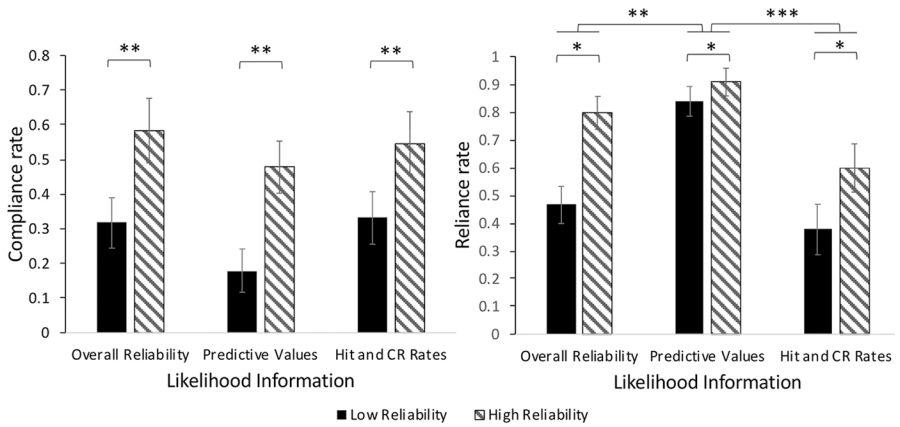


Figure 3. Compliance with and reliance on automated threat detector. *Difference is significant at the .05 level (two-tailed). **Difference is significant at the .01 level. ***Difference is significant at the .001 level. CR = correct rejection.

effect of likelihood information was non-significant.

Reliance. Automation reliability, $F(1, 56) = 5.905$, $p = .018$, and likelihood information, $F(2, 56) = 10.752$, $p < .001$, significantly affected reliance rate. Higher automation reliability led to higher reliance. Moreover, providing participants with predictive values led to higher reliance on the automated threat detector, compared with the overall success likelihood condition ($p = .009$) and the hit/CR rates condition ($p < .001$). There was also a significant two-way interaction between automation reliability and likelihood information, $F(2, 56) = 4.807$, $p = .012$. When automation reliability was low, participants relied on the automated threat detector the most when they were informed of the predictive values (predictive values > overall success likelihood: $p < .001$; predictive values > hit/CR rates: $p < .001$). As reliability increased, the reliance rate was significantly higher when participants were provided with predictive values relative to the hit/CR rates ($p = .004$).

Performance

Detection performance. As depicted in Figure 4, participants detected threats more accurately, $F(1, 56) = 9.702$, $p = .003$, and faster, $F(1, 56) = 8.659$, $p = .005$, and gained higher scores, $F(1, 56) = 14.633$, $p < .001$, when automation reliability increased. However, the effect of likelihood information was not significant.

Tracking performance. As shown in Figure 5, there were significant main effects of automation reliability, $F(1, 56) = 4.37$, $p = .041$, and likelihood information, $F(2, 56) = 5.381$, $p = .007$, on tracking score. Post hoc analysis indicated that when participants were presented with hit/CR rates, they had the lowest tracking score (hit/CR rates < predictive values: $p = .038$; hit/CR rates < overall success likelihood: $p = .011$).

In addition, there was a significant effect of likelihood information, $F(2, 56) = 4.311$, $p = .018$, on RMSE. When participants were presented with hit/CR rates, they had a higher RMSE (hit/CR rates > overall success likelihood: $p = .019$). The main effect of automation reliability on RMSE was not significant.

Combined performance. The main effects of automation reliability, $F(1, 56) = 10.744$, $p = .002$, and likelihood information, $F(2, 56) = 6.293$, $p = .003$, were significant (Figure 6). Participants obtained higher combined scores as automation reliability increased. There was also a difference among the three types of likelihood information. Post hoc analysis revealed that participants informed of overall success likelihood or predictive values, instead of the hit/CR rates, had higher total scores (overall success likelihood > hit/CR rates: $p = .008$; predictive values > hit/CR rates: $p = .014$).

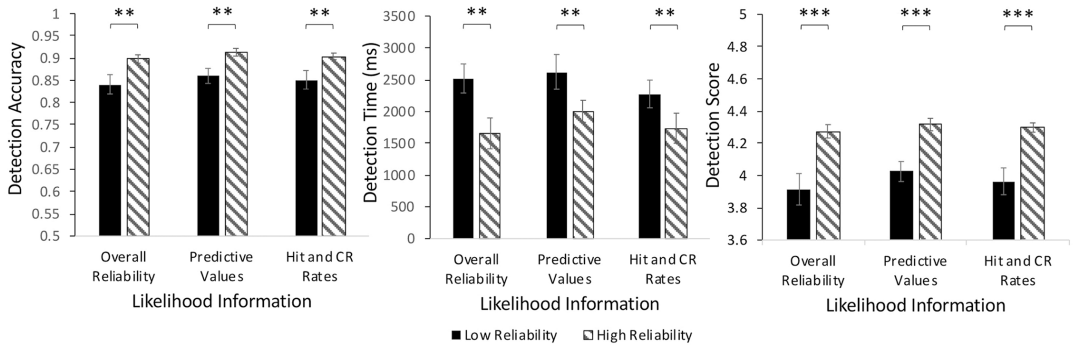


Figure 4. Detection task performance. CR = correct rejection.

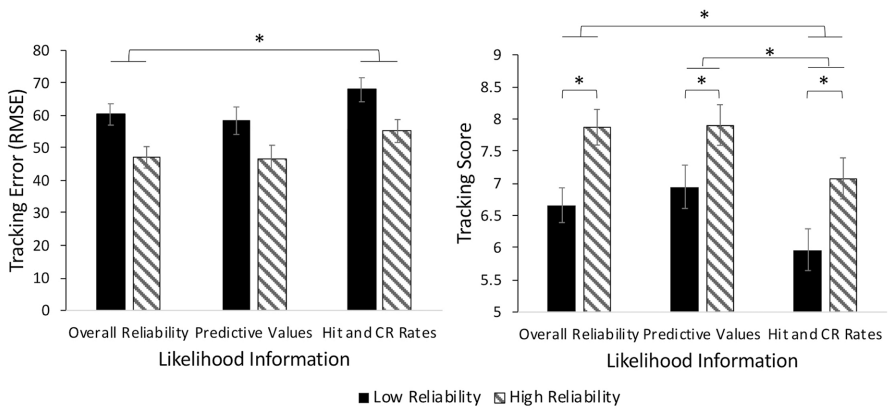


Figure 5. Tracking task performance. CR = correct rejection.

DISCUSSION

In the present study, we predicted that there would be significant differences in participants' trust, dependence, and dual-task performance when presented with different types of likelihood information. In particular, disclosing hit/CR rate would be the least beneficial in fostering proper trust and compliance and reliance behaviors and would lead to the worst task performance. Revealing the predictive values, in contrast, would be the most beneficial. We discuss how the results support our prediction.

Trust in Automation

Our results indicate a nonsignificant difference on trust between the three types of likelihood information. The lack of significance might have been due to two reasons. First, the sensitivity of a unidimensional trust scale might not be as high as that of a multidimensional scale. A unidimensional scale has the advantage of easy implementation. However, it might not be able to capture the different dimensions underlying the concept of trust compared with the multidimensional scales. Two widely used multidimensional scales (Jian, Bisantz, & Drury, 2000; Muir & Moray, 1996) have 12 and 7 questions, respectively. Second, the reliability of the threat detector was consistent across the different types of likelihood information. Therefore, the judgment of trust might be largely based on the true performance of automated detector instead of the likelihood information presentation. Further research is needed to systematically examine potential differences between unidimensional and multidimensional trust scales.

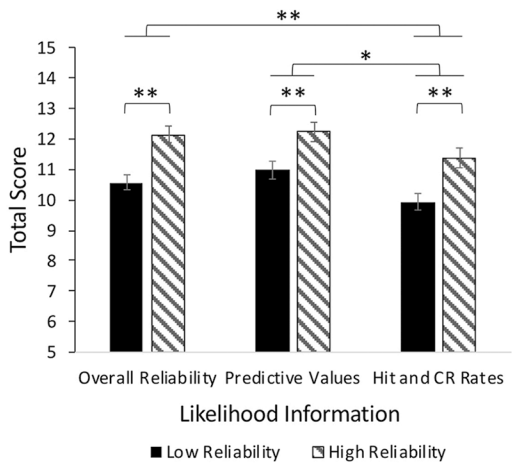


Figure 6. Total task performance. CR = correct rejection.

Compliance and Reliance Behaviors

We found a significant difference in reliance and a nonsignificant difference in compliance between the three types of likelihood information. Disclosing the predictive values led to higher and more appropriate reliance, compared with the overall success likelihood condition and the hit/CR rates condition. We argue that the predictive values can be considered as the gold standard of optimal behaviors. The negative predictive value, Pr(No signal|No alert) = x%, means that when the automation is silent, there is x% chance that a site is clear. Therefore, probablistically speaking, if the threat detector is silent for 100 cases, the human operator only needs to check 100 – x sites in person.

In our study, the negative predictive value was 97% for the high reliability automation and 92% for the low reliability automation. Therefore, a rational strategy for the human operator is to cross-check only a small number of sites and to allocate more resource on the tracking task. When presented with negative predictive values, participants' reliance rates were 90.8% and 83.8%, respectively (see Figure 3 and Table 2), which were fairly close to the optional values of 97% and 92%. When informed of the overall likelihood, the observed reliance values were 79.7% and 46.5%, respectively, further away from the optimal values; when presented with the hit/CR rates, the observed reliance values were 59.8% and 37.7%, respectively, furthest away from the optimal values. In the present study, the base rate was set to be 30%. In real life, base rates of critical events are usually much lower (Parasuraman & Riley, 1997). With a lower base rate, most of the time, the automated decision aid would be silent, and the benefit of presenting the predictive values would be further enhanced, as the predictive values promote proper reliance behaviors.

We failed to find a significance in participant's compliance behaviors. This lack of significance might have resulted from participants' strategies between the detection and the tracking tasks. The positive predictive values were 78% for the high reliability automation and 54% for the low reliability automation. However, across all the likelihood conditions, the compliance rates were considerably lower than the optimal values (see Figure 3 and Table 2). This suggests that participants cross-checked the detection display much more frequently than they should have done. This is further supported by our observation: Participants mentioned that the tracking task was fairly boring and they preferred to cross-checking the detection display even if the strategy was not optimal. The unnecessary cross-checking behaviors allowed the participants to detect threats that the automated detector failed to recognize and contributed to a similar performance in the detection task.

Performance

Our results indicate a significant difference in tracking task and nonsignificant difference in detection task. The tracking performances in the predictive value condition and the overall likelihood condition were better than that in the hit/CR rate condition. Such results are attributable to participants' reliance and compliance behaviors. When presented with hit/CR rates, participants' reliance behaviors were the least optimal, which means they cross-checked much more frequently than they should have done. Every time a cross-checking was performed, participants could not access the tracking display, hurting the tracking performance. In addition, as mentioned before, the similar compliance behaviors resulted in the similar performance in the detection task.

TABLE 2: Mean and Standard Error of Dependent Variables in Each Condition

	Low Reliability, $c = -0.25$, $d' = 1.5$			High Reliability, $c = -0.25$, $d' = 3$		
	Overall Success Likelihood	Predictive Values	Hit/CR Rates	Overall Success Likelihood	Predictive Values	Hit/CR Rates
Trust	53.15 ± 4.98	53.21 ± 3.70	58.82 ± 4.22	76.53 ± 2.79	71.60 ± 2.57	69.55 ± 54.92
Compliance (%)	31.67 ± 7.27	17.92 ± 6.24	33.13 ± 7.58	58.33 ± 9.30	47.78 ± 7.51	54.72 ± 9.03
Reliance (%)	46.54 ± 6.67	83.85 ± 5.36	37.69 ± 9.07	79.69 ± 5.92	90.78 ± 5.01	59.84 ± 8.68
Detection time (ms)	2,518.90 ± 228.54	2,623.29 ± 272.74	2,275.49 ± 217.99	1,655.10 ± 242.21	2,002.61 ± 173.16	1,736.20 ± 236.60
Detection accuracy (%)	84.10 ± 2.17	86.00 ± 1.68	85.10 ± 2.11	89.80 ± 0.93	91.30 ± 0.87	90.40 ± 0.72
Detection score	3.91 ± 0.10	4.03 ± 0.06	3.96 ± 0.08	4.27 ± 0.04	4.33 ± 0.04	4.30 ± 0.03
Tracking error	60.30 ± 3.28	58.37 ± 4.23	67.89 ± 3.70	47.10 ± 3.31	46.78 ± 4.04	55.20 ± 3.53
Tracking score	6.66 ± 0.27	6.95 ± 0.34	5.97 ± 0.32	7.87 ± 0.28	7.91 ± 0.32	7.08 ± 0.32
Total score	10.58 ± 0.24	10.97 ± 0.29	9.93 ± 0.27	12.15 ± 0.27	12.22 ± 0.32	11.38 ± 0.32

Note. CR = correct rejection.

The observed pattern on tracking and detection performance suggests that the automated threat detector was largely used as a tool for attention management in multitask environments, benefiting the continuous unaided task (i.e., the tracking task), rather than a tool directly benefiting the aided task (i.e., the detection task). The result supports the findings of Wiczorek and Manzey (2014).

In addition, we also observed a difference in the combined task performance. Disclosing predictive values and overall likelihood information led to higher combined performance than the hit/CR rates condition. We note the importance of obtaining an explicit payoff structure with the same unit of measurement. Most of the prior literature did not report the combined task performance, largely because different tasks were measured in different units and a combined task performance score was impossible to obtain.

At last, consistent with findings from previous studies, our results showed that as the automated threat detector became more reliable, participants' trust in and dependence on the threat detector increased, and their dual-task performance improved (Neyedli et al., 2011; Walliser et al., 2016; Wang et al., 2009).

CONCLUSION

Although disclosing likelihood information has been proposed as a design solution to promote proper trust and dependence, and to enhance human-automation team performance, prior studies showed mixed results (Bagheri & Jamieson, 2004; Dzindolet et al., 2002; Fletcher et al., 2017; Walliser et al., 2016; Wang et al., 2009). The goal of this study was to experimentally examine the effects of presenting different types of likelihood information. Based on the framework of SDT, we categorized likelihood information calculated in prior literature into three types: overall success likelihood, predictive values, and hit/CR rates.

The present study offered a framework to summarize existing literature pertaining to disclosing likelihood information. Our results showed that not all likelihood information is equally useful. Simply presenting the hit/CR rates should be avoided. Our findings can be applied to a wide array of domains such as urban search and rescue (USAR), medical diagnosis, and TSA, where the hit/CR rates are often readily available but not the predictive values and overall likelihood information. Hit/CR rates, also known as sensitivity and specificity (Altman & Bland, 1994), are referred to as the diagnostic information (Please note that the sensitivity as hit rate is different from the sensitivity d' in SDT). Often, the diagnostic information is more accessible to people. For instance, physicians are often provided with the diagnostic information when a new test is introduced: The HIV test is 99% accurate—if a patient is infected by HIV, there is 99% chance the test will show a positive result; if a patient is healthy, there is 99% chance the test will show a negative result.

Efforts should be made to clarify the meanings of different types of likelihood information when an automated decision aid is introduced. Prior research indicates that people could be confused about predictive values and hit/CR rates (Bar-Hillel, 1980). In real life, base rates of critical events are usually very low (Parasuraman & Riley, 1997). With a lower base rate, for instance, 1% in the airport security officer example, the discrepancies between the predictive values and the hit/CR rates would be even larger. Misattributing hit/CR rates as predictive values would lead to more detrimental outcomes.

The findings should be viewed in light of the following limitations. First, consistent with prior research, we did not provide participants with the base rate and they had to estimate it by themselves. Future study can present base rate to participants and examine whether people can use hit/CR rates more appropriately. Base rate can also be manipulated in further research to examine the effects of likelihood information when base rate is extremely low. Second, we used probabilities instead of natural frequencies to present likelihood information. Previous studies have shown that reasoning with natural frequencies results in more accurate inference (Gigerenzer & Hoffrage, 1995; Hoffrage, Hafenbrädl, & Bouquet, 2015; Mandel, 2014). A future study could compare the differences of presenting probabilities and natural frequencies. Third, the criterion c in this study was set to be liberal, which led to more FAs than misses. Future

studies should examine the effects of likelihood information with different d' and c.

APPENDIX

Pilot Study

We conducted a pilot study to create a scoring system for the experiment. In the experimental task, participants performed the tracking task and the detection task simultaneously. Participants were required to make a trade-off decision on which task to perform at any time, that is, if they decided to check the four images, they would probably earn more points in the detection task but fewer points in the tracking task, and vice versa.

Therefore, a payoff structure has to be determined to eliminate potential bias toward either the tracking or the detection task by ensuring that the potential gain in one task is approximately equal to the opportunity cost in the other task. To determine the parameters of the scoring system, first we set the tracking task score on a scale from 0 to 10, based on the distance of the green circle from the center of the crosshair. Next, we defined the detection task score as a function of the detection accuracy and time, that is, $a + b \times (\text{time} / 10{,}000 \text{ ms})$. To determine a and b, a total of 10 participants between the age of 19 and 23 years participated in the pilot study. They performed a tracking task only block and a combined task block, each with 50 trials, with a 5-min break in between. In the combined task block, participants performed both tasks and were instructed that two tasks were equally important. They could optimize their performance by minimizing the distance between the green circle and the center of the display, and by detecting the threats as accurately and as quickly as possible. The block order was counterbalanced. One participant's data were removed from data analysis due to his significantly poor performance on the tracking task. The results showed that when doing both tasks concurrently, participants lost on average a score of 3.7 points on their tracking tasks (tracking task only block: $M = 8.8$, $SD = 1.2$; combined task block: $M = 5.1$, $SD = 1.1$). We then varied a and b to make sure they would gain approximately 3.7 points with a similar SD from the detection task. As a result, $5 - 5 \times (\text{detection time}/10{,}000 \text{ ms})$ was set to be the scoring scheme of the detection task. In each combined task trial, it is possible to obtain a maximum score of 15 points, 10 points from the tracking task, and 5 points from the detection task.

ACKNOWLEDGMENTS

The authors would like to thank Kevin Li and Benjamin Pinzone for programming the simulator.

KEY POINTS

- We proposed a framework to summarize existing literature pertaining to disclosing likelihood information and classified the calculation of likelihood information into three categories: overall likelihood value, predictive values, and hit and correct rejection (CR) rates.
- Human operators informed of the overall likelihood value or the predictive values, rather than the hit and CR rates, relied on the decision aid more appropriately.
- Human operators informed of the overall likelihood value or the predictive values, rather than the hit and CR rates, performed better on the tracking task and obtained higher combined task scores.
- As automation reliability increased, trust, compliance, reliance, and performance increased accordingly.

REFERENCES

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *British Medical Journal*, *308*(6943), 1552.

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. In *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics* (pp. 212–217). Netherlands: The Hague.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233.

Comstock, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behaviour research* (Tech. Rep. No. 104174). Washington, DC: National Aeronautics and Space Administration.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*, 564–572.

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference,

anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*, 79–94.

Fletcher, K. I., Bartlett, M. L., Cockshell, S. J., & McCarley, J. S. (2017). Visualizing probability of detection to aid sonar operator performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, pp. 302–306). Santa Monica, CA: Human Factors and Ergonomics Society.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Goodie, A. S., & Fantino, E. (1996, March). Learning to commit or avoid the base-rate error. *Nature*, *380*, 247–249. doi:10 .1038/380247a0

Hoffrage, U., Hafenbrädl, S., & Bouquet. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*, Article 642.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*, 53–71.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80.

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide. Mahwah, NJ: Lawrence Erlbaum.

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers in Psychology*, *5*, Article 1144.

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011, November). Understanding the effect of workload on automation use for younger and older adults. *Human Factors*, *53*, 672–686.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*, 656–665.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*, 429–460.

Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human Factors*, *53*, 338–355.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253.

Petersen, L., Robert, L., Yang, X. J., & Tilbury, D. (2019). Situational awareness, driver's trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Automated Vehicles*, *2*(2), 129–141.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors*, *30*, 445–459.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 401–409.

Walliser, J. C., de Visser, E. J., & Shaw, T. H. (2016). Application of a system-wide trust strategy when supervising multiple autonomous agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60. pp. 133–137). Santa Monica, CA: Human Factors and Ergonomics Society.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*, *51*, 281–291.

Wiczorek, R., & Manzey, D. (2014). Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors*, *56*, 1209–1221.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)* (pp. 408–416). Vienna, Austria.

Na Du is a PhD student in the Department of Industrial and Operations Engineering at the University of Michigan Ann Arbor. She completed a BS in psychology in Zhejiang University, China.

Kevin Y. Huang is an undergraduate student studying industrial and operations engineering at the University of Michigan Ann Arbor.

X. Jessie Yang is an assistant professor in the Department of Industrial and Operations Engineering and an affiliated faculty at the Robotics Institute, University of Michigan Ann Arbor. She obtained her PhD in mechanical and aerospace engineering (Human Factors) from Nanyang Technological University Singapore in 2014.

Date received: November 30, 2018
Date accepted: June 14, 2019

Effects of Initial Starting Distance and Gap Characteristics on Children's and Young Adults' Velocity Regulation When Intercepting Moving Gaps

Hyun Chae Chung, **Gyoojae Choi**, and **Muhammad Azam**,
Kunsan National University, Republic of Korea

Objective: This study investigated how children and young adults regulate their velocity when crossing roads under varying traffic conditions.

Background: To cross roads safely, pedestrians must adapt their movements to the moving vehicles around them while tightly coupling their movement to visual information.

Method: Using an Oculus Rift, 16 children and 16 young adults walked on a treadmill and intercepted gaps between two simulated moving vehicles in an immersive virtual environment. We varied the participants' initial distance from the curb to the interception point, as well as gap characteristics, including gap size and vehicle size.

Results: Varying the initial distance led to systematic adjustments in participants' approach velocities. The inter-vehicle gap and the vehicle size affected the crossing position induced by the initial distance. However, participants did not systematically scale their positions according to the initial distance in narrow gap. Notably, children did not finely tune their movements when they approached wide gap from a closer distance or when they approached the large vehicle from closer distance.

Conclusion: Children were less precise in coupling their movements to the moving vehicle in complex traffic environments. In particular, large moving vehicles approaching at closer distances can pose risks when children cross roads.

Application: These findings suggest the need for an intervention program to improve children's skill in perceiving larger vehicles and timing their movements when crossing roads. We suggest using an interactive virtual reality system to practice this skill.

Keywords: gap crossing, coupling, perception-action, virtual reality, speed

Address correspondence to Hyun Chae Chung, Department of Sport and Exercise Sciences, Kunsan National University, 63 Myrong dong, JB, Gunsan 54150, Republic of Korea; e-mail: hcx@kunsan.ac.kr.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 1002–1018
DOI: 10.1177/0018720819867501

Korea has one of the highest child traffic death rates worldwide at 0.9 per 100,000 children under the age of 14 years (Traffic Accident Analysis System, 2016). Furthermore, 50.7% of child pedestrian deaths occur in traffic accidents while crossing a road (Traffic Accident Analysis System, 2016). These statistics highlight the importance of understanding children's road-crossing behavior.

Gap crossing is an interceptive behavior that requires pedestrians to move in relation to the open space between two moving vehicles (Chihak et al., 2010; Louveton, Montagne, Berthelon, & Bootsma, 2012). Accordingly, gap crossing involves not only perceiving oncoming vehicles, but also controlling one's movement in relation to moving traffic. To cross a road successfully, individuals must time their movements to those of moving vehicles; this requires precise coupling of actions with visual information.

Researchers investigating goal-directed behavior control found that children's gap-crossing behavior was less finely tuned than that of adults (Chihak, Grechkin, Kearney, Cremer, & Plumert, 2014; Chihak et al., 2010; O'Neal et al., 2018; Plumert, Kearney, & Cremer, 2004; Plumert, Kearney, Cremer, Recker, & Strutt, 2011; Te Velde, Van der Kamp, & Savelsbergh, 2008). For example, Chihak et al. (2010) reported that when 10- and 12-year-old cyclists crossed 12 intersections, the children experienced difficulty timing their actions with the movement of car-sized blocks, and they have less time to spare after clearing approaching traffic blocks. This inefficiency is likely due to the children's lack of skill in coordinating their locomotion with moving traffic. From late childhood through early

adolescence, children undergo developmental changes in their skill to coordinate movements to match moving objects (Grechkin, Chihak, Cremer, Kearney, & Plumert, 2013; O'Neal et al., 2018; Savelsbergh, Rosengren, Van der Kamp, & Verheul, 2003). O'Neal et al. (2018) investigated how 6-, 8-, 10-, 12-, and 14-year-old children and adults pedestrians perceive and act on dynamic affordances when crossing roads. They found that 12-year-old children exhibited poorer timing of gap behind lead vehicle (LV) in the gap than 14-year-olds and adults. However, research has not yet explored how 12-year-old children regulate their velocities in various changing environments. Thus, we further examined the road-crossing behavior of 12-year-old children in various environments.

Given that children lack the skill to coordinate their movements with moving traffic, previous studies have extensively investigated velocity control during gap crossing. These studies characterized 10- to 12-year-old children's velocity regulation as an overcorrection in speed when they moved (Chihak et al., 2010; Chihak et al., 2014) and found that children have less time to spare than adults (Grechkin et al., 2013; Plumert et al., 2004; Plumert et al., 2011). In a study of pedestrian behavior, Te Velde et al. (2008) investigated age-related differences in child pedestrian road-crossing behavior by moving a doll between two toy vehicles to simulate crossing a road. They found that 5- to 7-year-old children reached the required velocity to avoid colliding with the second vehicle later than preadolescent children and adults. However, this study did not involve actual walking. People who actually cross a road can better judge the time gap than people who only make a verbal decision to cross (Oudejans, Michaels, Van Dort, & Frissen, 1996). In total, these results indicated that children are less skillful than young adults in scaling their movements based on visual information.

We investigated children's velocity regulation of children's road crossing behaviors by incorporating an actual crossing in a virtual reality environment. Crossing the gap in a changing environment is a complex task in which it is necessary to scale locomotion in relation to the moving traffic. Studies on gap-crossing behavior in cyclists and drivers (Dewing, Duley, & Hancock, 1993; Louveton, Bootsma, Guerin, Berthelon, & Montagne, 2012; Louveton, Montagne et al., 2012) indicated that crossing environment influences crossing behaviors.

The gap, a moving object that must be intercepted, is an important feature of the traffic environment that should affect crossing behavior. Louveton, Montagne et al. (2012) studied global (gap-related) and local (vehicle-related) gap manipulation and found that the inter-vehicle gap between LVs and trailing vehicles (TVs) contributed to changes in drivers' regulation of road-crossing speed, leading them to cross earlier in a wider traffic gap. Similarly, studies of locomotion indicated that people must adjust their walking speed to maintain a constant relationship with the moving objects to be intercepted, thereby yielding a successful interception (see Chardenon, Montagne, Laurent, & Bootsma, 2004). The information related by the spatial-temporal characteristics of the intercepted object specifies how the actor can move. In the gap-crossing context, changing the gap size should affect how pedestrians regulate their speed.

Another aspect of the traffic environment that may affect pedestrian crossing behavior is vehicle size. Hancock, Caird, Shekhar, and Vercruyssen (1991) found that drivers chose to turn left across traffic more frequently in front of smaller oncoming vehicles. Mathieu, Bootsma, Berthelon, and Montagne (2017) studied the effects of vehicle size and type on an intersection-crossing driving task and found that participants crossed the intersection slightly slower when they encountered a double-sized vehicle rather than a normal-sized vehicle at the final stage of the approach. Thus, these studies imply that dynamic gap characteristics may influence velocity adjustment and its effect on crossing position. Functionally appropriate gap-crossing behavior requires pedestrians to scale their movements based on dynamic information about moving vehicles. This demands a precise coupling of his or her action with the visual information. As such, vehicle size should affect his or her velocity regulation.

We compared how children and young adults regulate their velocities when crossing a street in

a virtual reality environment. The participants' task was to cross the gap between two moving vehicles on a virtual road. First, we systematically manipulated the participants' initial distance from the curb to the interception point to create an offset within the gap. If participants maintained a constant walking speed, varying initial distance should have led to early, on time, or late arrival at the center of gap. This offset allowed us to determine participants' velocity adjustments when approaching the interception point, similar to the paradigm used in previous research (Chihak et al., 2010; Louveton, Bootsma et al., 2012). Second, we manipulated gap characteristics by varying gap and vehicle size to investigate whether these changes affect children's velocity control.

We hypothesized that changing the initial starting distance would affect children and young adults' velocity adjustment when approaching the interception point, leading them to cross at different positions within the gap at the moment of interception and enter the gap at different times. More specifically, we expected that children would adjust their velocity less adeptly than young adults. In addition, we expected that manipulating gap and vehicle size would lead children to deviate more from the center of gap and take longer to enter the gap.

METHODS

Participants

We recruited 16 children (mean age = 12.18 years, *SD* = 0.83) and 16 young adults (mean age = 22.75 years, *SD* = 2.56) with normal or corrected-to-normal vision. All participants volunteered. Two young adults experienced motion sickness during the experiment and were replaced to match the group. Participants signed written consent forms, and the Kunsan National University Research Board approved the experimental procedure. The minimum sample size to achieve power for our study was 24 within the given parameters (effect size = 0.2, $\alpha = 0.05$, power = 0.9).

Apparatus and Virtual Environments

We conducted the experiment using a walking simulator consisting of a customized treadmill (0.67 m wide × 1.26 m long × 1.10 m high), an Oculus Rift (DK1, US), and a PC (3.30 GHz with 8.00 GB RAM, Figure 1). Participants walked on the treadmill using their own locomotive skill; the treadmill was equipped with a handrail for their safety. Participants also wore a hook and loop belt secured to the back of the treadmill to decrease vertical and lateral movements, and four magnetic counters on a spinning roller recorded the participants' displacement.

We presented the virtual environment using an Oculus Rift (1,280 × 800 pixels) that produced 3-D stereoscopic images. The visual scene changed in accordance with participants' walking speed.

Experimental Setup and Procedure

The virtual street consisted of a two-lane road (3.5 m per lane), trees, and a building-lined skyline, as well as a general street view of the road (see Figure 1). We manipulated three experimental variables: participants' initial distance, gap size, and vehicle size. Walking speed is approximately 1.0 to 1.67 m/s for most adults, 1.17 m/s for children aged 6 to 12 years, and 1.22 m/s for teenagers (Waters & Mulroy, 1999). Thus, we set the initial distance from the curb to the interception point assuming that participants would walk at an average speed of 1.1 m/s. Under such conditions, participants would successfully cross the gap further ahead of, near, and further away from the center of gap for near (3.5 m), intermediate (4.5 m), and far (5.5 m) initial distances, respectively.

The gap, treated as an entity (see Chihak et al., 2010; Louveton, Montagne et al., 2012), was defined as the space between the rear bumper of the LV and the front bumper of the TV. The arrival of the gap center was set to 4 s (around 33.2 m) from the interception point. We established the gap size using two vehicles moving at a constant speed of 8.33 m/s with an intervehicular distance of 24.9 m (temporal gap of 3 s) or 33.2 m (temporal gap of 4 s). These gap sizes were chosen because O'Neal et al.'s (2018) study of pedestrian road crossing in a virtual environment showed that a 4-s crossing gap is comfortable, whereas a 3-s gap is tight but crossable. We varied vehicle size based on previous research (Mathieu et al., 2017) indicating that

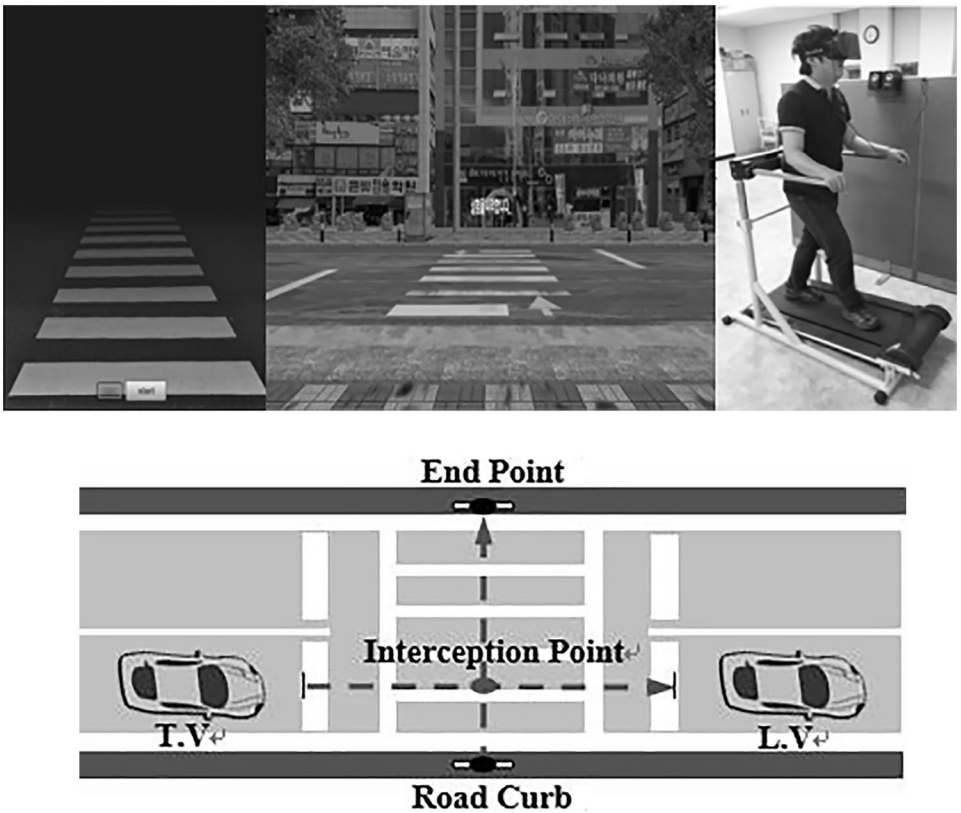


Figure 1. A black-and-white cartoon image of the crosswalk (top left), the street view (top middle), the walking simulator (top right), and a schematic view of the virtual road (bottom). TV represents the trailing vehicle, and LV represents the lead vehicle.

vehicle size affects participants' crossing behavior. The simulation presented either two white sedans (1.5 m wide, 3.5 m long) or two orange buses (2.4 m wide, 11 m long). The vehicles appeared on the left side of the road in the near lane. No vehicles occupied the far lane.

The participants' task was to safely cross the gap between two vehicles traveling at a constant speed of 8.33 m/s (around 30 km/h) and walk until arriving on the other side of the virtual road. At the beginning of the trial, participants viewed a black-and-white cartoon image of the virtual crosswalk to calibrate the street view. At the verbal *ready* signal, participants prepared to cross; at the *go* signal, the experimenter pressed a button to start the vehicles' motion and participants were required to look left immediately, visualize the oncoming vehicles, and cross the road if the gap was safe to cross. Participants completed six practice trials intended to familiarize them with the task and the virtual environment. These consisted of two free-walking trials without the head-mounted display, two trials without any vehicles, and two trials in which the vehicles moved at a constant speed of 25 km/h with a 5-s inter-vehicle gap. Following the practice trials, participants performed the task twice under each set of experimental conditions (3 initial distances × 2 gap sizes × 2 vehicle sizes), resulting in a total of 24 trials.

The word *success, collision*, or *failure* appeared at the end of each trial. The word *success* appeared if a participant successfully crossed the gap and reached the other side of the road. *Collision* and *failure* appeared if a participant collided with the vehicle or missed the gap, respectively. After each trial, the experimenter restarted the simulation by pushing a button. Presentation order was

counter balanced across participants. We repeated the trial twice because Plumert et al. (2011) reported that short-term changes occurred after specific road-crossing experiences. If participants experienced motion sickness, we ceased data collection and excluded their data from the analysis.

Data Analysis

We evaluated participants' crossing behavior via (a) each participant's position and velocity profile while approaching the interception point, (b) gap entry time, and (c) position within the gap at the moment of interception.

To examine the participants' velocity regulation changes in position and velocity as the participants approached the interception point were averaged into 1-s intervals (–3.5 s, –2.5 s, –1.5 s, and –.5 s) counting backward from the participants' arrival at the interception point (e.g., Chihak et al., 2014; Louveton, Montagne et al., 2012). We examined participants' positions and velocities to evaluate their speed adjustment and its instantaneous effect on position within the gap during approach.

We calculated mean gap entry time for each trial to evaluate how participants adjusted their movements within the available time. We examined gap entry time to evaluate participants' temporal distance from the LV. Smaller values indicated that participants crossed the gap closer to the LV with more time to spare between them self and the TV.

We evaluated the participants' deviation from the gap center at the moment of interception as the time of interception (TOI). TOI can be defined as the temporal distance between the time at which participants crossed the interception point and the time at which the center of gap arrived at the participants' crossing line. We evaluated TOI as the instantaneous effect of speed adjustment on participants' position within the gap, and we average TOI for each trial. Negative TOI indicates participants crossed before the center of gap, and positive TOI indicates participants crossed after the center of gap. Multiplying this value by vehicle speed (8.33 m/s) yields the actual position within the gap (in meters).

We analyzed position and velocity data using initial distance (near, intermediate, far) × gap size (3 s, 4 s) × vehicle size (car, bus) × time (3.5 s, 2.5 s, 1.5 s, 0.5 s) repeated measures analysis of variance (ANOVA), with initial distance, gap size, vehicle size, and time as within-factor variables. The timing data were analyzed using initial distance (near, intermediate, far) × gap size (3 s, 4 s) × vehicle size (car, bus) repeated measures ANOVA, with initial distance, gap size, and vehicle size as within-factor variables. The partial eta squared (η_p^2) was used to estimate effect size. A least square mean was used for all pairwise post hoc comparisons, and *p*-values were adjusted using a Bonferroni correction to decrease type I errors. SAS software (version 9.4) was used for the data analysis.

RESULTS

Across all participants, the success rate was 98.95% for children and 99.48% for young adults. We analyzed only the data for successful trials to access the participants' crossing behaviors and time of crossing. We do not discuss the results of the frequency analysis here because it is beyond the scope of this paper.

We tested our hypothesis that changing the initial distance would affect the participants' approach position and velocity, and that manipulating gap characteristics would affect children's and young adults' approach positions and the velocity profiles induced by the initial distance.

Approach Position

Young adults. Young adults adjusted their crossing positions according to initial distance while crossing the gap (see Figure 2 for an example of an individual young adult). As the initial distance became further away, young adults crossed the gap closer to the TV.

A repeated measures ANOVA of approaching position showed significant main effects of initial distance, $F(2, 30) = 1{,}289.10$, $p < .0001$, $\eta_p^2 = .99$, and gap size, $F(1, 15) = 9.60$, $p < .007$, $\eta_p^2 = .39$. Young adults' mean position to the interception point increased with the initial distance. In addition, young adults' mean position to the interception point was greater for the 4-s gap than for the 3-s gap (Table 1).

The initial distance × time interaction was also significant, $F(6, 90) = 230.26$, $p < .0001$,

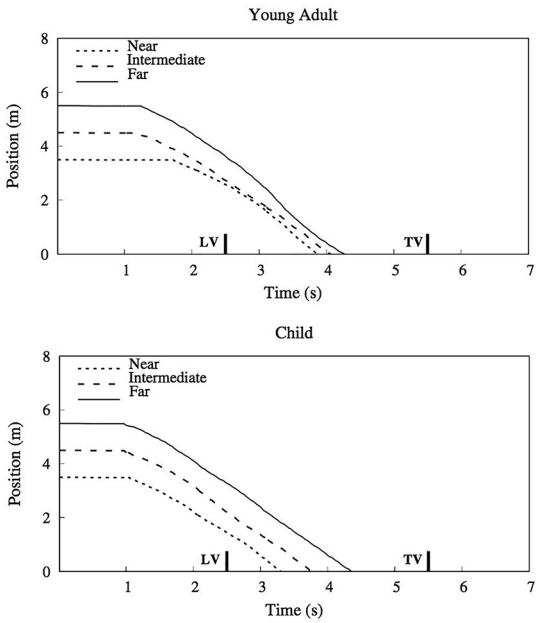


Figure 2. The sample trajectories of a young adult and a child in relation to the LV and TV during a successful gap crossing. TV represents the trailing vehicle and LV represents the lead vehicle.

$\eta_p^2 = .94$. A simple effects test showed a significant effect of time for the near initial distance, $F(3, 45) = 1{,}313.07$, $p < .0001$, $\eta_p^2 = .99$; the intermediate initial distance, $F(3, 45) = 4{,}472.97$, $p < .0001$, $\eta_p^2 = .99$; and the far initial distance, $F(3, 45) = 8{,}779.54$, $p < .0001$, $\eta_p^2 = .99$. Post hoc comparisons revealed that young adults' crossing position as determined by initial distance significantly decreased from 3.5 to 0.5 s (all *ps* < .0001) before reaching the interception point (Figure 3). Young adults' mean position to interception point decreased as they approached it. In addition, the mean position increased with initial distance.

Children. Children adjusted their crossing positions according to the initial distance while crossing the gap (see Figure 2 for an example of an individual child). Similar to young adults, children crossed the gap closer to the TV as the initial distance increased.

A repeated measures ANOVA on approaching position showed significant main effects of initial distance, $F(2, 30) = 2{,}059.46$, $p < .0001$, = .99; gap size, $F(1, 15) = 11.70$, $p < .004$, $\eta_p^2 = .44$; and vehicle size, $F(1, 15) = 10.60$, $p < .005$, $\eta_p^2 = .41$. The children's mean position to the interception point was greater for the far initial distance compared with the near initial distance. In addition, the children's mean position to the interception point was greater for the 4-s gap than for the 3-s gap. It was also greater when crossing between cars than when crossing between the buses (Table 1).

The initial distance × time interaction was also significant, $F(6, 90) = 412.28$, $p < .0001$, $\eta_p^2 = .96$. A simple effects test showed a significant effect of time for near initial distance, $F(3, 45) = 3{,}861.11$, $p < .0001$, $\eta_p^2 = .99$; intermediate initial distance, $F(3, 45) = 7{,}115.29$, $p < .0001$, $\eta_p^2 = .99$; and far initial distance, $F(3, 45) = 14{,}490.3$, $p < .0001$, $\eta_p^2 = .99$. Post hoc comparisons revealed that children's crossing positions induced by initial distance decreased significantly from 3.5 to 0.5 s (all $p < .0001$) before reaching the interception point (Figure 3). Children's mean position to interception point decreased as they approached it. In addition, the mean position increased with initial distance.

Velocity Profiles

As we expected, participants adjusted their velocities differently according to the initial distances while approaching the interception point. We observed that initial distance influenced participants' velocity patterns when they encountered different gap and vehicle sizes.

Young adults. A repeated-measures ANOVA on velocity profiles showed significant main effects of initial distance, $F(2, 30) = 29.62$, $p < .0001$, $\eta_p^2 = .66$, and gap size, $F(1, 15) = 10.93$, $p < .005$, $\eta_p^2 = .42$. Young adults crossed the gap faster as the initial distance became further away. They also crossed the 4-s gap faster than the 3-s gap (Table 1).

The initial distance × time interaction was also significant, $F(6, 90) = 11.88$, $p < .0001$, $\eta_p^2 = .44$. A simple effects test showed a significant effect of time for near initial distance, $F(3, 45) = 140.34$, $p < .0001$, $\eta_p^2 = .90$; intermediate initial distance, $F(3, 45) = 29.93$, $p < .0001$, $\eta_p^2 = .67$; and far initial distance, $F(3, 45) = 184.46$, $p < .0001$, $\eta_p^2 = .93$. Post hoc comparisons showed that for the near initial distance, young adults' velocity significantly decreased from

TABLE 1: Mean Position, Velocity, Gap Entry Time, and Time of Interception (*SD*) as a Function of Initial Distance, Gap Size, and Vehicle Size for Children and Young Adults

	Position (m)		Velocity (m/s)		Gap Entry Time (s)		Time of Interception (s)	
	Children	Young Adults	Children	Young Adults	Children	Young Adults	Children	Young Adults
Initial distance								
Near	2.40 (1.03)	2.50 (1.02)	0.92 (0.54)	0.99 (0.69)	3.48 (0.33)	3.35 (0.38)	–0.26 (0.32)	–0.42 (0.39)
Intermediate	2.87 (1.38)	3.02 (1.36)	1.14 (0.55)	1.23 (0.99)	3.65 (0.36)	3.54 (0.37)	–0.08 (0.36)	–0.24 (0.37)
Far	3.25 (1.65)	3.47 (1.68)	1.33 (0.53)	1.39 (0.71)	3.95 (0.29)	3.75 (0.37)	0.21 (0.32)	–0.01 (0.40)
Gap size								
3 s	2.81 (1.40)	2.98 (1.43)	1.10 (0.58)	1.14 (0.67)	3.82 (0.32)	3.67 (0.35)	0.09 (0.33)	–0.08 (0.37)
4 s	2.86 (1.42)	3.02 (1.45)	1.16 (0.55)	1.28 (0.95)	3.57 (0.39)	3.42 (0.42)	–0.18 (0.39)	–0.36 (0.43)
Vehicle size								
Car	2.86 (1.42)	3.00 (1.45)	1.14 (0.57)	1.23 (0.88)	3.66 (0.42)	3.45 (0.40)	–0.04 (0.43)	–0.28 (0.42)
Bus	2.81 (1.41)	3.00 (1.43)	1.11 (0.56)	1.18 (0.75)	3.72 (0.33)	3.64 (0.40)	–0.05 (0.33)	–0.16 (0.41)

Note. Near = 3.5 m initial distance; intermediate = 4.5 m initial distance; far = 5.5 m initial distance.

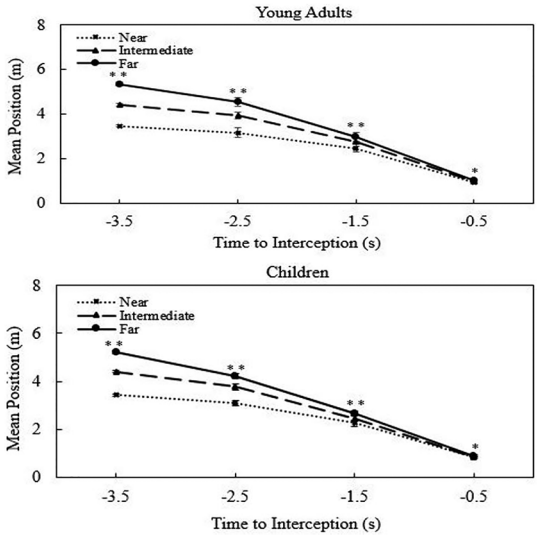


Figure 3. Young adults and children's mean approach positions for each initial distance (near, intermediate, and far) as a function of time before reaching the interception point. The participants' position while approaching the interception point was averaged into 1-s intervals (–3.5 s, –2.5 s, –1.5 s, and –.5 s), counting backward from the interception point. In the figure, asterisks represent statistically significant inter-mean differences for initial distances at each time point. One asterisk represents one inter-mean difference, and two asterisks represent two or more inter-mean differences. Error bars indicate standard deviations.

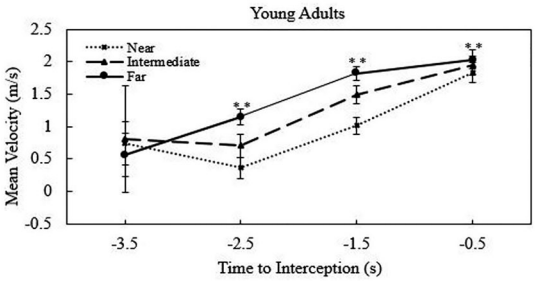


Figure 4. Young adults' mean velocity for each initial distance (near, intermediate and far) as a function of time before reaching the interception point. The approaching velocity was averaged into 1-s intervals (–3.5 s, –2.5 s, –1.5 s, and –.5 s) counting backward from the interception point. In the figure, asterisks represent statistically significant inter-mean differences for initial distances at each time point. One asterisk represents one inter-mean difference, and two asterisks represent two or more inter-mean differences. Error bars indicate standard deviations. Error bars indicate standard deviations.

3.5 s to 2.5 s ($p < .0001$) and increased from 2.5 to 0.5 s ($p < .0001$) before reaching the interception point. For the intermediate initial distance, young adults' velocity significantly increased from 2.5 to 1.5 s ($p < .0001$) and from 1.5 to 0.5 s ($p < .02$) before reaching the interception point. For the far initial distance, young adults' velocity significantly increased from 3.5 to 1.5 s ($p < .0001$) and from 1.5 to 0.5 s ($p < .03$) before reaching the interception point (Figure 4). For the most part, young adults increased their speed throughout the approach, but for the near initial distance, they decreased their speed at the beginning of the approach.

In addition, there was a significant interaction effect of gap size × time, $F(3, 45) = 7.95$, $p < .0002$, $\eta_p^2 = .35$. A simple effects test showed a significant effect of time for the 3-s gap, $F(3, 45) = 268.31$, $p < .0001$, $\eta_p^2 = .95$; and for the 4-s gap, $F(3, 45) = 47.80$, $p < .0001$, $\eta_p^2 = .76$. Post hoc comparisons showed that for the 3-s gap, young adults' velocity significantly increased from 3.5 to 0.5 s ($p < .0001$) before reaching the interception point. For the 4-s gap, young adults' velocity significantly increased from 2.5 to 1.5 s ($p < .0001$) and from 1.5 to 0.5 s ($p < .002$) before reaching the interception point (Table 2). Young adults did not speed up at the beginning of approach (3.5–2.5 s) for the 4-s gap, but they increased their speed during the rest of approach to the interception point. In addition, young adults crossed the 4-s gap faster than the 3-s gap during the beginning ($p < .003$) and middle (2.5–1.5 s; $p < .04$) approach phases.

Children. A repeated-measures ANOVA on velocity profile showed significant main effects of initial distance, $F(2, 30) = 207.32$, $p < .0001$, $\eta_p^2 = .93$, and gap size, $F(1, 15) = 13.44$, $p < .002$, $\eta_p^2 = .47$. Children crossed the gap faster as the initial distance became further away. They also crossed the 4-s gap faster than the 3-s gap (see Table 1).

Initial distance × time interaction was also significant, $F(6, 90) = 53.51$, $p < .0001$, $\eta_p^2 = .78$. This interaction effect was captured by the three-way interaction. In addition, the gap size ×

TABLE 2: Mean Velocities (*SD*) of Young Adults and Children for Gap Size as a Function of Time Before Reaching the Interception Point

	Young Adults				Children			
	–3.5 s	–2.5 s	–1.5 s	–0.5 s	–3.5 s	–2.5 s	–1.5 s	–0.5 s
3-s (m/s)	0.48 (0.30)	0.78 (0.45)	1.38 (0.44)	1.92 (0.29)	0.42 (0.29)	0.90 (0.43)	1.36 (0.33)	1.70 (0.19)
4-s (m/s)	0.93 (1.51)	0.70 (0.44)	1.51 (0.38)	1.95 (0.32)	0.57 (0.33)	0.89 (0.43)	1.47 (0.29)	1.70 (0.20)
p value	*		*		*		*	

Note. Asterisk indicates statistically significant inter-mean differences for gap size at each time point.

time interaction was significant, $F(3, 45) = 5.98$, $p < .002$, $\eta_p^2 = .29$. A simple effects test showed a significant effect of time for the 3-s gap, $F(3, 45) = 266.81$, $p < .0001$, $\eta_p^2 = .95$, and for the 4-s gap, $F(3, 45) = 235.24$, $p < .0001$, $\eta_p^2 = .94$. Post hoc comparisons indicated that for both gaps, children's velocity significantly increased from 3.5 to 0.5 s (all, $p < .0001$) before reaching the interception point (see Table 2). Children consistently increased their speed throughout the approach for both gap sizes. In addition, they crossed the 4-s gap faster than the 3-s gap during the beginning ($p < .0006$) and middle ($p < .003$) approach phases.

The vehicle size × initial distance × time interaction was significant, $F(6, 90) = 2.12$, $p < .05$, $\eta_p^2 = .12$. Further analysis revealed that, between the cars, the initial distance × time interaction was significant, $F(6, 90) = 33.55$, $p < .0001$, $\eta_p^2 = .69$. A simple effects test showed a significant effect of time for near initial distance, $F(3, 45) = 132.54$, $p < .0001$, $\eta_p^2 = .90$; intermediate initial distance, $F(3, 45) = 173.83$, $p < .0001$, $\eta_p^2 = .92$; and far initial distance, $F(3, 45) = 272.78$, $p < .0001$, $\eta_p^2 = .95$. Post hoc comparisons showed that when participants crossed between the cars, for near initial distance, children's velocity significantly decreased from 3.5 to 2.5 s ($p < .0002$), but it increased from 2.5 to 0.5 s ($p < .0001$) before reaching the interception point. For intermediate initial distance, children's velocity significantly increased from 3.5 to 1.5 s ($p < .0001$) and from 1.5 to 0.5 s ($p < .01$) before reaching the interception point. For the far initial distance, children's velocity significantly increased from 3.5 to 1.5 s ($p < .0001$) before reaching the interception point (Figure 5). For the most part, children increased their speed throughout their approaches, but their speed decreased at the beginning of the approach for the near initial distance, when they crossed between the cars.

When participants crossed between the buses, the initial distance × time interaction was also significant, $F(6, 90) = 18.70$, $p < .0001$, $\eta_p^2 = .55$. A simple effects test showed a significant effect of time for the near initial distance, $F(3, 45) = 124.41$, $p < .0001$, $\eta_p^2 = .89$; intermediate initial distance, $F(3, 45) = 132.79$, $p < .0001$, $\eta_p^2 = .90$; and far initial distance, $F(3, 45) = 331.16$, $p < .0001$, $\eta_p^2 = .96$. Post hoc comparisons showed that, for the near initial distance, children's velocity significantly increased from 2.5 to 0.5 s ($p < .0001$) before reaching the interception point. For the intermediate initial distance, children's velocity increased from 3.5 to 0.5 s ($p < .0001$) before reaching the interception point. For the far initial distance, children also crossed the gap significantly faster from 3.5 to 1.5 s ($p < .0001$) and from 1.5 to 0.5 s ($p < .03$) before reaching the interception point (Figure 5). When children crossed between the buses, their speed neither increased nor decreased at the beginning of their approach for the near initial distance.

Gap Entry Time

We tested our hypothesis that the initial distance and manipulated gap characteristics would affect participants' gap entry time.

Young adults. A repeated-measures ANOVA on gap entry time showed significant main effects of initial distance, $F(2, 30) = 44.60$, $p < .0001$, $\eta_p^2 = .75$; gap size, $F(1, 15) = 57.80$, $p < .0001$, $\eta_p^2 = .79$; and vehicle size, $F(1,15) = 27.63$, $p < .0001$, $\eta_p^2 = .65$. Young adults crossed

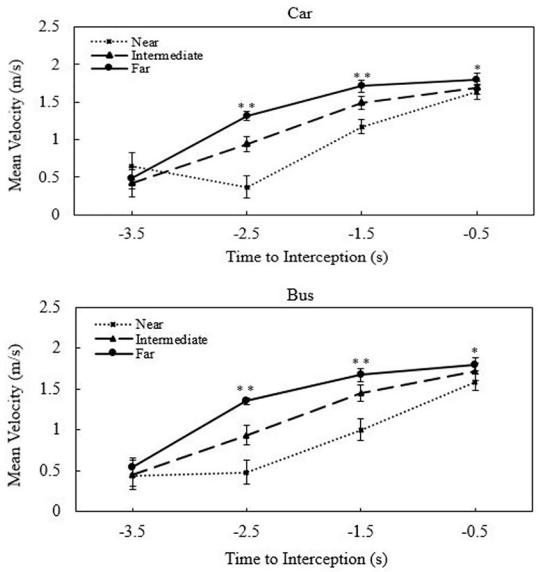


Figure 5. Children's mean velocity profiles before reaching the interception point for each vehicle and for each initial distance (near, intermediate, or far) as a function of time. The approach velocity was averaged into 1-s intervals (–3.5 s, –2.5 s, –1.5 s, and –.5 s), counting backward from the interception point. In the figure, asterisks represent statistically significant inter-mean differences for initial distances at each time point. One asterisk represents one inter-mean difference, and two asterisks represent two or more inter-mean differences. Error bars indicate standard deviations.

the gap earlier and closer to the LV when initial distance decreased. They also crossed the gap earlier and closer to the LV for the 4-s gap compared with the 3-s gap, as well as when crossing between the cars compared with crossing between the buses (see Table 1).

The gap size × initial distance interaction was also significant, $F(2, 30) = 5.53$, $p < .009$, $\eta_p^2 = .27$. A simple effects test showed a significant effect of initial distance for the 3-s gap, $F(2, 30) = 8.93$, $p < .0009$, $\eta_p^2 = .37$, and for the 4-s gap, $F(2, 30) = 37.13$, $p < .0001$, $\eta_p^2 = .71$. Post hoc comparisons showed that, for the 3-s gap, young adults crossed the gap later when the initial distance changed from intermediate to far ($p < .01$). For the 4-s gap, young adults crossed the gap later when the initial distance changed from near to intermediate ($p < .0001$) and from intermediate to far ($p < .002$, Table 3). Young adults crossed the gap later and closer to the TV as the initial distance increased for the 4-s gap, but for 3-s gap, they crossed the gap at similar times for near and intermediate initial distance.

Children. A repeated-measures ANOVA on gap entry time showed significant main effects of initial distance, $F(2, 30) = 67.94$, $p < .0001$, $\eta_p^2 = .82$, and gap size, $F(1, 15) = 68.26$, $p < .0001$, $\eta_p^2 = .82$. Children crossed the gap earlier and closer to the LV when initial distances decreased. They also crossed the gap earlier and closer to the LV for the 4-s gap than for the 3-s gap (Table 1).

The gap size × initial distance interaction was significant, $F(2, 30) = 3.97$, $p < .03$, $\eta_p^2 = .21$. A simple effects test showed a significant effect of initial distance for the 3-s gap, $F(2, 30) = 12.81$, $p < .0001$, $\eta_p^2 = .46$, and for the 4-s gap, $F(2, 30) = 50.58$, $p < .0001$, $\eta_p^2 = .77$. Post hoc comparisons showed that, for the 3-s gap, children crossed the gap later for the far initial distance than for the intermediate initial distance ($p < .01$). For the 4-s gap, children crossed the gap later for the intermediate initial distance compared with the near initial distance ($p < .007$) and for the far initial distance compared with intermediate initial distance ($p < .0001$, Table 3). Similar to young adults, children crossed the gap later and closer to the TV as the initial distance increased for the 4-s gap, but for 3-s gap, they crossed the gap at similar times for near and intermediate initial distance.

The vehicle size × initial distance interaction was significant, $F(2, 30) = 18.40$, $p < .0001$, $\eta_p^2 = .55$. A simple effects test showed a significant effect of initial distance between the cars, $F(2, 30) = 64.81$, $p < .0001$, $\eta_p^2 = .81$, and between the buses, $F(2, 30) = 6.63$, $p < .004$, $\eta_p^2 = 31$. Post hoc comparisons revealed that between the cars, children crossed the gap later when the initial distance increased from near to far (near: $M = 3.32$ s, $SD = 0.29$; intermediate: $M = 3.67$ s, $SD = 0.36$; far: $M = 4.00$ s, $SD = 0.28$; $p < .0001$). Between the buses, children's gap entry time was not significantly different when comparing near and intermediate initial distances ($p = 1$), but it significantly increased for the far initial distance compared with the

TABLE 3: Young Adults and Children's Mean Gap Entry Time (*SD*) for Different Gap Sizes as a Function of Initial Distance

	Young Adults			Children		
	Near	Intermediate	Far	Near	Intermediate	Far
3-s (s)	3.55 (0.28)	3.63 (0.35)	3.83 (0.37)	3.66 (0.31)	3.80 (0.31)	4.01 (0.27)
4-s (s)	3.15 (0.38)	3.45 (0.37)	3.67 (0.36)	3.31 (0.26)	3.50 (0.36)	3.90 (0.30)
p value	*	*	*	*	*	*

Note. Asterisk indicates statistically significant inter-mean differences for gap size at each initial distance.

intermediate initial distance (intermediate: $M = 3.63$ s, $SD = 0.36$; far: $M = 3.89$ s, $SD = .28$; $p < .008$). Thus, when they crossed between the cars, children crossed the gap earlier and closer to the LV as the initial distance increased, but when they crossed between the buses, they crossed the gap at similar times for near and intermediate initial distance.

The vehicle size × gap size interaction was significant, $F(1, 15) = 5.50$, $p < .03$, $\eta_p^2 = .27$. A simple effects test showed a significant effect of gap size between the cars, $F(1, 15) = 5.67$, $p < .03$, $\eta_p^2 = .27$, and between the buses, $F(1, 15) = 36.15$, $p < .0001$, $\eta_p^2 = .71$. Post hoc comparisons showed that, when crossing between the cars, children crossed the gap earlier and closer to the LV for the 4-s gap ($M = 3.57$ s, $SD = 0.06$) than for the 3-s gap ($M = 3.75$ s, $SD = 0.06$, $p < .03$). When crossing between the buses, children also crossed the gap earlier and closer to the LV for the 4-s gap ($M = 3.55$ s, $SD = 0.04$) than for the 3-s gap ($M = 3.89$ s, $SD = 0.04$, $p < .0001$). In addition, for both vehicle sizes, children crossed the gap earlier and closer to the LV when crossing 4-s gap than 3-s gap.

Time of Interception

We tested our hypothesis that changing the initial distance and manipulating the gap characteristics would cause deviation in participants' crossing positions from the center of the gap at the moment of interception. Velocity adjustment while approaching the interception point led participants to cross the gap closer to either the LV or the TV even though participants crossed the gap near its center. Systematic velocity regulation led the participants to arrive at the gap early or late depending on their initial distances.

Young adults. A repeated measures ANOVA on TOI showed significant main effects of initial distance, $F(2, 30) = 44.12$, $p < .0001$, $\eta_p^2 = .75$; gap size, $F(1, 15) = 65.66$, $p < .0001$, $\eta_p^2 = .81$; and vehicle size, $F(1, 15) = 12.5$, $p < .003$, $\eta_p^2 = .45$. Young adults crossed the gap furthest ahead of the gap center for the near initial distance, further ahead of the gap center for the intermediate initial distance, and near the gap center for the far initial distance. In addition, young adults crossed the gap further ahead of the gap center for the 4-s gap than for the 3-s gap (Table 1).

The gap size × initial distance interaction was significant, $F(2, 30) = 5.39$, $p < .01$, $\eta_p^2 = .26$. A simple effects test showed a significant effect of initial distance for the 3-s gap, $F(2, 30) = 11.07$, $p < .0003$, $\eta_p^2 = .43$, and for the 4-s gap, $F(2, 30) = 37.98$, $p < .0001$, $\eta_p^2 = 72$. Post hoc comparisons showed that, for the 3-s gap, young adults crossed the gap closer to the gap center as the initial distance increased from intermediate to far ($p < .002$). For the 4-s gap, young adults crossed the gap significantly closer to the gap center as the initial distance increased from near to intermediate ($p < .0001$) and intermediate to far ($p < .003$, Figure 6). For the 4-s gap, young adults' deviation from the gap center was significantly larger as the initial distance became further away, but for the 3-s gap, they crossed at similar positions relative to the gap center for near and intermediate initial distances.

Children. A repeated measures ANOVA on TOI showed significant main effects of initial distance, $F(2, 30) = 63.98$, $p < .0001$, $\eta_p^2 = .81$, and gap size, $F(1, 15) = 69.81$, $p < .0001$, $\eta_p^2 = .82$. Children crossed the gap further ahead of the gap center at the near initial distance, near

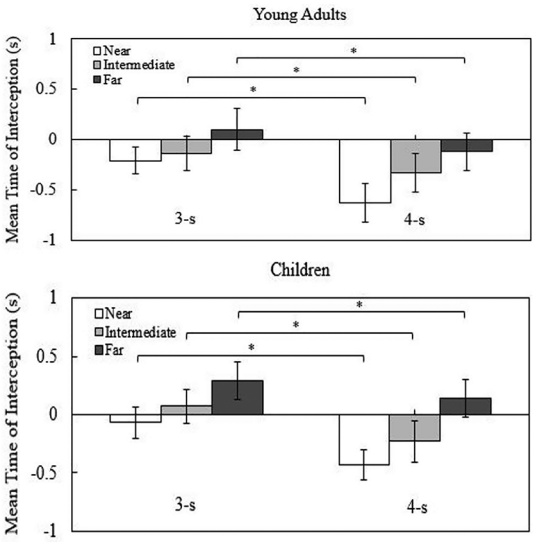


Figure 6. Young adults and children's mean time of interception (TOI) for each initial distance (near, intermediate, or far) as a function of gap size (3-s, 4-s). TOI refers to the temporal distance relative to the gap center, such that 0.2 s would refer to around 1.6 m when vehicle speed is 30 km/h (8.3 m/s). In the figure, asterisks represent statistically significant inter-mean differences for gap size at each initial distance. One asterisk represents one inter-mean difference, and two asterisks represent two or more inter-mean differences. Error bars indicate standard deviations.

to the gap center at the intermediate initial distance, and further away from the gap center at the far initial distance. In addition, children crossed the gap further ahead of the gap center for the 4-s gap than for the 3-s gap (see Table 1).

The gap size × initial distance interaction was significant, $F(2, 30) = 3.48$, $p < .04$, $\eta_p^2 = .19$. A simple effects test showed a significant effect of initial distance for the 3-s gap, $F(2, 30) = 14.74$, $p < .0001$, $\eta_p^2 = .50$, and for the 4-s gap, $F(2, 30) = 43.34$, $p < .0001$, $\eta_p^2 = .74$. Post hoc comparisons showed that, for the 3-s gap, children crossed the gap further away from the gap center as the initial distance increased from intermediate to far ($p < .004$). For the 4-s gap, children crossed the gap significantly further away from the gap center when comparing near to intermediate ($p < .008$) and intermediate to far initial distances ($p < .0001$, see Figure 6). For the 4-s gap, children crossed the gap systematically further away from the gap center as the initial distance increased. However, for the 3-s gap, children crossed at similar position relative to the gap center for the near and intermediate initial distances.

The vehicle size × initial distance interaction was significant, $F(2, 30) = 18.13$, $p < .0001$, $\eta_p^2 = .55$. A simple effects test showed a significant effect of initial distance between cars, $F(2, 30) = 62.30$, $p < .0001$, $\eta_p^2 = .81$, and between buses, $F(2, 30) = 6.15$, $p < .005$, $\eta_p^2 = .30$. Post hoc comparisons showed that between the cars, children crossed the gap systematically further ahead of the center of the gap for near, the gap center for intermediate, and further away for far initial distances, respectively (all $p < .0001$). However, between the buses, children crossed the gap further ahead of the gap center as the initial distance increased from intermediate to far ($p < .01$, Figure 7). Thus, children crossed at similar positions relative to the gap center for near and intermediate initial distances when they crossed between the buses.

The vehicle size × gap size interaction was significant, $F(1, 15) = 4.26$, $p < .05$, $\eta_p^2 = .22$. A simple effects test showed a significant effect of gap size between the cars, $F(1, 15) = 7.42$, $p < .02$, $\eta_p^2 = .33$, and between the buses, $F(1, 15) = 35.93$, $p < .001$, $\eta_p^2 = .71$. Post hoc comparisons showed that when crossing between the cars, children crossed the gap significantly further ahead of the gap center for the 4-s gap ($M = -0.14$, $SD = 0.07$) than for the 3-s gap ($M = 0.06$ s, $SD = 0.07$, $p < .01$). When crossing between the buses, children also crossed the gap significantly further ahead of the gap center for the 4-s gap ($M = -0.12$ s, $SD = .04$) than for the 3-s gap ($M = 0.12$ s, $SD = .04$, $p < .0001$). Children crossed the gap further ahead of the gap center for the 4-s gap than for the 3-s gap for both vehicles.

DISCUSSION

We designed this study to evaluate how children and young adults adjust their crossing behaviors in response to moving traffic gaps in changing traffic environments. As expected, the participants' systematic positions and velocity

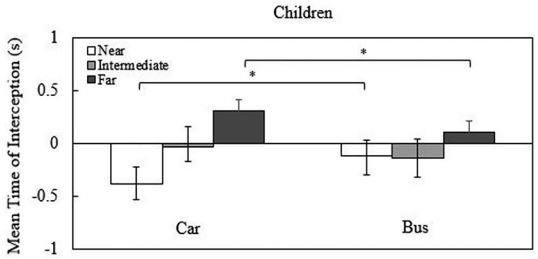


Figure 7. Children's mean time of interception (TOI) for each initial distance (near, intermediate or far) as a function of vehicle size (car, bus). TOI refers to the temporal distance relative to the gap center, such that 0.2 s refers to around 1.6 m when vehicle speed is 30 km/h (8.3 m/s). In the figure, asterisks represent statistically significant inter-mean differences for gap size at each initial distance. One asterisk represents one inter-mean difference, and two asterisks represent two or more inter-mean differences. Error bars indicate standard deviations.

adjustments led them to cross at different positions within the gap. Varying gap and vehicle size affected children's and young adults' gap-crossing behavior differently. Young adults and children crossed the gap faster and closer to the LV for the wide (4-s gap) gap than for the narrow (3-s gap) gap. However, participants did not fine-tune their movements according to the initial distances when they crossed the narrow gap. In particular, children did not adjust their movements in relation to moving vehicles when they approached the wide gap from closer distances. Furthermore, children did not adjust their velocities relative to the initial distances when they approached the large vehicle from closer distances. We discuss these findings in more detail in terms of initial distance and gap characteristics below.

Effects of Initial Distance

A systematic change in the initial distances affected children's and young adults' velocity adjustments. The participants' approach positions and velocity profiles while approaching the interception point varied according to the initial distances. Participants adjusted their velocities while approaching the interception point instead of making last-moment adjustments. The results confirmed previous findings about the crossing behaviors of drivers and cyclists (Chihak et al., 2010; Louveton, Montagne et al., 2012; Mathieu et al., 2017), which showed that the last moment of acceleration did not fully compensate for the initial offset. In our study, participants also sped up at the last moment of interception for all initial distances, but the crossing-point discrepancy resulting from the initial-distance variation persisted until the last moment of interception. Although deviations from the gap center in the gap-crossing times systematically varied (around a 0.2-s difference for each initial distance) depending on the initial distances, the participants crossed the gap near its center.

Children and young adults made functional adjustments to their velocities to achieve their goals. For example, participants decreased their velocities at the beginning of the trial in the near initial distance condition, but they maintained and increased their velocities while approaching the interception point in the intermediate initial distance condition, and they continuously increased their velocities in the far initial distance condition. This resulted in similar position profiles for young adults and children, although the children's crossing positions within the gap shifted slightly at the last moment compared with those of the young adults. Evidently, the children and young adults regulated and timed their movements based on the initial distances according to their capabilities (Oudejans et al., 1996). Specifically, children passed near the center of the gap in the intermediate initial distance condition, but young adults passed near the center of the gap in the far initial distance condition. This systematically adaptive crossing behavior reflects the coupling of perception-action in road crossing (Gibson, 1979).

Effects of Gap Characteristics

Gap size manipulation affected participants' gap-crossing behaviors. Young adults and children crossed the gap faster and closer to the LV when they crossed the wide gap than when crossing the narrow gap as shown in previous studies (Louveton, Bootsma et al., 2012; Louveton, Montagne et al., 2012). In our experimental setup, the LV in the 4-s gap was closer

to the interception point compared with the LV in the 3-s gap. Thus, this result reflects safe crossing behavior as Louveton, Bootsma et al. (2012) suggested. Furthermore, gap size affected the crossing position induced by initial distance. For the wide gap, young adults and children adjusted their crossing positions systematically depending on the initial distances. However, participants' crossing positions did not systemically vary according to the initial distances when they crossed the narrow gap (see Figure 6). When they crossed the narrow gap in the near initial distance condition, participants took longer to initiate movements and did not compensate for their longer initiation times with increased speed. Narrow gaps therefore appear to pose challenges for young adults and children. Participants did not adjust their movements according to the initial distances if they had less available time to cross.

Specifically, the children's velocity profiles displayed continuous speeding up when they approached the interception point for both gap sizes. However, for the wide gap, young adults maintained and somewhat decreased their speeds at the beginning of the trial but sped up during the remainder of it. When young adults entered the wide gap, they realized they had more time available before arriving at the TV and thus lowered their speeds to adapt. However, children did not adjust their walking speeds according to the available crossing time (see Lee, Young, & McLaughlin, 1984). Children seemed to control their movements based on the LV movement without considering the TV when they approached the wide gap from closer distances. These results also aligned with previous findings regarding children's poor coordination of movement with moving vehicles (Chihak et al., 2010; O'Neal et al., 2018), and they imply that 12-year-old children have not yet developed the skill of synchronizing their movements in relation to moving objects when they face time constraints.

Our results clearly showed the effect of vehicle size on participants' timing and crossing behaviors. Noticeably, young adults crossed the gap further ahead of the gap center when facing a small vehicle than when facing a large vehicle. In addition, the children's positions were farther away from the gap center between the buses than between the cars. The results are novel in that they reveal the effect of vehicle size on intercepting pedestrian gap-crossing behavior. Our results do not align with earlier studies' findings on the effect of size-distance prediction on perceptual judgment—that is, that individuals perceive larger objects as closer when compared with smaller objects (Caird & Hancock, 1994; DeLucia, 1991; DeLucia & Warren, 1994). The effects of size on perceptual judgment are not compatible with our observed crossing behavior as Mathieu et al. (2017) suggested.

Vehicle size interacted with initial distance to influence children's crossing behaviors. The children's crossing positions did not deviate based on the initial distances when they crossed in front of the large vehicle. However, they displayed a systematic deviation from the gap center depending on the initial distances when they crossed in front of the small vehicle. The result supports Grechkin et al.'s (2013) findings that children did not coordinate their movements according to the visual information as skillfully as young adults did. In front of a large vehicle, children crossed the gap less far ahead from the gap center than expected for the near initial distance condition. The result reflected that children may overestimate the TV's arrival time and may therefore attempt to cross more slowly in front of a large vehicle. The result indicates that children might ignore the speed-related information of large moving vehicles and rely exclusively on distance information. This can lead children to fail to estimate the TV's arrival time. This interpretation was further supported by a longer than expected gap entry time for the near initial distance condition. This result indicates that children took longer to initiate their movements in front of a large vehicle in the near initial distance condition. Specifically, children did not adjust their velocities according to the initial distances at the beginning when they crossed between the buses (see Figure 6). Our velocity analysis revealed that children did not speed up at the beginning of trial in the near initial distance condition. This indicates that children did not compensate for their longer initiation times by increasing their velocities when they faced a

large vehicle approaching at closer distances. The results imply that children face problems in controlling their velocities and in timing their movements in complex traffic environments as a previous study (O'Neal et al., 2018) suggested.

Limitations and Future Research

The safety margin referred to the difference between the time a pedestrian crossed the traffic and the time the TV's front bumper arrived at the pedestrian's crossing point (Chu & Baltes, 2001). The successes and failures reported in this study may not generalize to real-world situations due to the lack of a safety margin. In this study, we considered a trial to be successful if the participant crossed between the vehicles and made it to the other side of the road without colliding with a vehicle. Thus, we did not account for a safety margin. Narrow escapes can be important issues to consider for collision prediction. Although we did not set up safety margins, the TOIs of those participants who crossed the gap closest from the TV and LV were at 0.93 s and 1.3 s, respectively, equivalent to distances of around 8 m and 10 m, respectively. This suggests that participants who crossed successfully did so near the gap center. Although this did not lead to close calls, future research addressing safety margins remains important.

Another limitation of our study is that we did not control for participants' heights and stride lengths. How fast an actor can move is specified by the perceived properties of the environment in relation to the perceiver's biomechanical dimensions and action capability (Fajen, 2013; Warren, 1984). Our results revealed potential evidence of the effects of various body sizes on crossing positions. However, considering physical variables, such as height and stride length, might yield different results.

CONCLUSION

In conclusion, varying initial distance, manipulating gap and vehicle size strongly and systematically influenced young adults' and children's gap-crossing behaviors. In addition, our findings clearly showed that children may experience difficulty coordinating their movements with visual information when they approach a large vehicle from closer distances and if they have time constraints, such as crossing narrow gaps and approaching inter-vehicle gaps from closer distances. Our findings could provide the first evidence of the clear effect of vehicle size on the crossing behaviors of children and young adults in various traffic environments. In addition, our study contributes to the understanding of children's crossing behaviors in relation to temporal and spatial gap characteristics in a paradigm that is highly ecologically valid. It is noteworthy that 12-year-old children are still undergoing developmental changes related to precisely coupling their movements in relation to moving objects in complex dynamic environments. Children must develop a tight link between perception and action to scale their movements in relation to moving objects in complex situations. Children need to learn the use of perceptual information and movement timing in interception actions as they physically grow and as their motor skills become refined.

Our results underscore the need for a training program that teaches children to synchronize themselves with moving vehicles in real-world traffic scenarios. An important practical application is the development of an intervention program that focuses on improving children's skill to control their velocities in dynamic traffic environments. Experience with various environmental crossing actions, including various vehicle sizes with various initial crossing distances, should be considered to reduce risk behavior by improving children's skill to link perception and action. An interactive virtual reality system is a promising tool for fine-tuning children's perceptions and actions and for linking their actions to the time available for crossing while allowing them to walk actively in a virtual environment. Future research should focus on the mechanisms underlying the control of children's crossing behaviors.

ACKNOWLEDGMENT

The Korea Institute funded this work for Advancement of Technology and Ministry of Trade, Industry, and Energy [grant number 10044775].

KEY POINTS

- We investigated children and young adults' velocity regulation while intercepting moving gap.
- Participants adjusted their approach to the interception based on initial distance.
- Children did not precisely adjust their movements to the moving vehicles when children approached the inter-vehicle gap from the closer distance.
- Children did not time their movement according to the initial distance when they approached large moving vehicles from closer distance.

ORCID iDS

Hyun Chae Chung https://orcid.org/0000-0002-6322-8120

Muhammad Azam https://orcid.org/0000-0002-1464-2595

REFERENCES

Caird, J. K., & Hancock, P. A. (1994). The perception of arrival time for different oncoming vehicles at an intersection. *Ecological Psychology*, *6*, 83–109.

Chardenon, A., Montagne, G., Laurent, M., & Bootsma, R. J. (2004). The perceptual control of goal-directed locomotion: A common control architecture for interception and navigation? *Experimental Brain Research*, *158*, 100–108.

Chihak, B. J., Grechkin, T. Y., Kearney, J. K., Cremer, J. F., & Plumert, J. M. (2014). How children and adults learn to intercept moving gaps. *Journal of Experimental Child Psychology*, *122*, 134–152.

Chihak, B. J., Plumert, J. M., Ziemer, C. J., Babu, S., Grechkin, T., Cremer, J. F., & Kearney, J. K. (2010). Synchronizing self and object movement: How child and adult cyclists intercept moving gaps in a virtual environment. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1535–1552.

Chu, X., & Baltes, M. R. (2001). *Pedestrian mid-block crossing difficulty* (Report No. NCTR 392-09). National Center for Transit Research for Florida Department of Transportation. Retrieved from https://www.nctr.usf.edu/pdf/PedMidblock.pdf

DeLucia, P. R. (1991). Pictorial and motion-based information for depth perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 738–748.

DeLucia, P. R., & Warren, R. (1994). Pictorial and motion-based depth information during active control of self-motion: Size-arrival effects on collision avoidance. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 783–798.

Dewing, W., Duley, J. A., & Hancock, P. A. (1993, October). *The role of vehicle type, velocity and gap size on driver left-turn decisions*. Paper presented at the 37th Annual Meeting of the Human Factor Society, Seattle, WA.

Fajen, R. B. (2013). Guiding locomotion in complex, dynamic environments. *Frontiers in Behavioral Neuroscience*, *7*: 85.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum.

Grechkin, T. Y., Chihak, B. J., Cremer, J. F., Kearney, J. K., & Plumert, J. M. (2013). Perceiving and acting on complex affordances: How children and adults bicycle across two lanes of opposing traffic. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 23–36.

Hancock, P. A., Caird, J. K., Shekhar, S., & Vercruyssen, M. (1991). Factors influencing drivers' left turn decisions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *35*, 1139–1143.

Lee, D. N., Young, D. S., & McLaughlin, C. M. (1984). A roadside simulation of road crossing for children. *Ergonomics*, *27*, 1271–1281.

Louveton, N., Bootsma, R. J., Guerin, P., Berthelon, C., & Montagne, G. (2012). Intersection crossing considered as intercepting a moving traffic gap: Effects of task and environmental constraints. *Acta Psychologica*, *141*, 287–294.

Louveton, N., Montagne, G., Berthelon, C., & Bootsma, R. J. (2012). Intercepting a moving traffic gap while avoiding collision with lead and trail vehicles: Gap-related and boundary-related influences on drivers' speed regulations during approach to an intersection. *Human Movement Science*, *31*, 1500–1516.

Mathieu, J., Bootsma, R. J., Berthelon, C., & Montagne, G. (2017). Judging arrival times of incoming traffic vehicles is not a prerequisite for safely crossing an intersection: Differential effects of vehicle size and type in passive judgment and active driving tasks. *Acta Psychologica*, *173*, 1–12.

O'Neal, E. E., Jiang, Y., Franzen, L. J., Rahimian, P., Yon, J. P., Kearney, J. K., & Plumert, J. M. (2018). Changes in perception-action tuning over long time scales: How children and adults perceive and act on dynamic affordances when crossing roads. *Journal of Experimental Psychology: Human Perception and Performance*, *44*, 18–26.

Oudejans, R. R., Michaels, C. F., Van Dort, B., & Frissen, E. J. P. (1996). To cross or not to cross: The effect of locomotion on street-crossing behavior. *Ecological Psychology*, *8*, 259–267.

Plumert, J. M., Kearney, J. K., & Cremer, J. F. (2004). Children's perception of gap affordances: Bicycling across traffic-filled intersections in an immersive virtual environment. *Child Development*, *75*, 1243–1253.

Plumert, J. M., Kearney, J. K., Cremer, J. F., Recker, K. M., & Strutt, J. (2011). Changes in children's perception-action tuning over short time scales: Bicycling across traffic-filled intersections in a virtual environment. *Journal of Experimental Child Psychology*, *108*, 322–337.

Savelsbergh, G. J. P., Rosengren, K. S., Van der Kamp, J., & Verheul, M. H. (2003). Catching action development. In Savelsbergh, G. J. P., Davids, K., Van der Kamp. J., & Bennett, S. J. (Eds.), *The development of movement co-ordination in children: Applications in the fields of ergonomics, health sciences and sport* (pp. 191–212). London, England: Taylor & Francis.

Te Velde, A. F., Van der Kamp, J., & Savelsbergh, G. J. P. (2008). Five- to twelve-year-old's control of movement velocity in a dynamic collision avoidance task. *British Journal of Developmental Psychology*, *26*, 33–50.

Traffic Accident Analysis System. (2016). *Traffic accident statistic report for Korea*. Retrieved from http://taas.koroad.or.kr/web/shp/sbm/initTfcacdStats.do

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 683–703.

Waters, R., & Mulroy, S. (1999). The energy expenditure of normal and pathologic gait. *Gait & Posture*, *9*, 207–231.

Hyun Chae Chung is the professor in the Department of Sport and Exercise Sciences at Kunsan National University, Republic of Korea. She earned her EdD in motor learning from the University of Georgia in 1995.

Gyoojae Choi is the professor in School of Mechanical and Automotive Engineering at Kunsan National University, Republic of Korea. He earned his PhD in mechanical engineering from the Korea Advanced Institute of Science and Technology in 2000.

Muhammad Azam is a PhD student in the Department of Sport and Exercise Sciences at Kunsan National University, Republic of Korea. He earned his MA in health and physical education from Sarhad University, Pakistan in 2013.

Date received: April 23, 2018
Date accepted: June 27, 2019

Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures

Anthony D. McDonald, **Thomas K. Ferris**, and **Tyler A. Wiener**,
Texas A&M University, College Station, USA

Objective: The objective of this study was to analyze a set of driver performance and physiological data using advanced machine learning approaches, including feature generation, to determine the best-performing algorithms for detecting driver distraction and predicting the source of distraction.

Background: Distracted driving is a causal factor in many vehicle crashes, often resulting in injuries and deaths. As mobile devices and in-vehicle information systems become more prevalent, the ability to detect and mitigate driver distraction becomes more important.

Method: This study trained 21 algorithms to identify when drivers were distracted by secondary cognitive and texting tasks. The algorithms included physiological and driving behavioral input processed with a comprehensive feature generation package, Time Series Feature Extraction based on Scalable Hypothesis tests.

Results: Results showed that a Random Forest algorithm, trained using only driving behavior measures and excluding driver physiological data, was the highest-performing algorithm for accurately classifying driver distraction. The most important input measures identified were lane offset, speed, and steering, whereas the most important feature types were standard deviation, quantiles, and nonlinear transforms.

Conclusion: This work suggests that distraction detection algorithms may be improved by considering ensemble machine learning algorithms that are trained with driving behavior measures and nonstandard features. In addition, the study presents several new indicators of distraction derived from speed and steering measures.

Application: Future development of distraction mitigation systems should focus on driver behavior–based algorithms that use complex feature generation techniques.

Keywords: distraction classification, cognitive distraction, machine learning, time-series feature generation, physiological measures

Address correspondence to Anthony D. McDonald, Texas A&M University, 4075 ETB, 3131 TAMU, College Station, TX 77843, USA; e-mail: mcdonald@tamu.edu.

HUMAN FACTORS
Vol. 62, No. 6, September 2020, pp. 1019–1035
DOI: 10.1177/0018720819856454
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Driver distraction is a major transportation safety problem. Official analysis of postcrash reports showed that drivers were found to be in a distracted state in at least 14% to 17% of vehicle crashes (National Center for Statistics and Analysis [NCSA], 2017), and other credible estimates show that distraction may be involved in as many as 68% of vehicle crashes (Dingus et al., 2016). The frequency and severity of distraction-affected crashes necessitate a comprehensive solution including legislation, training, and technology.

Although in-vehicle technologies are often a contributing factor in driver distraction, they can also be part of the solution. Distraction mitigation systems can algorithmically combine real-time input from vehicle and driver sensors to estimate the driver distraction, which can then be used to inform adaptive automation such as driver assist systems, alert and guide the driver when attentional reorientation is deemed necessary, and/or provide postdrive feedback to the driver (Kim, Chun, & Dey, 2015; Lee et al., 2013; Schwarz, Brown, Lee, Gaspar, & Kang, 2016; Smith, Witt, Bakowski, Leblanc, & Lee, 2009). Although most prior work in distraction estimation is based on linear or logistic modeling approaches, more recent work has investigated machine learning approaches to facilitate modeling highly nonlinear behavior that can be more sensitive to some types of distraction (e.g., Masood, Rai, Aggarwal, Doja, & Ahmad, 2018; Schwarz et al., 2016).

The challenge of determining when drivers are distracted can be approached using supervised machine learning models. Existing patterns of data associated with different types of distracted driving can be used to develop an algorithm capable of predicting future, unlabeled patterns (Dong, Hu, Uchimura, & Murayama, 2011;

Kotsiantis, 2007). The large body of research on driving distraction has documented how different types of distraction can affect vehicle control input (Dingus et al., 2016; Drews, Yazdani, Godfrey, Cooper, & Strayer, 2009; Engström, Markkula, Victor, & Merat, 2017; Feng et al., 2017; Horrey & Wickens, 2006; Strayer, Drews, & Johnston, 2003), driver head posture and eye-gaze (Lee et al., 2013; Schwarz et al., 2016; Tippey, Sivaraj, & Ferris, 2017), and physiological indicators of arousal in a driver's sympathetic nervous system, such as heart rate measures, galvanic skin response, or perinasal perspiration (Collet, Guillot, & Petit, 2010; Healey & Picard, 2005; Kim et al., 2015; Mehler, Reimer, Coughlin, & Dusek, 2009; Pavlidis et al., 2016; Reimer, Mehler, Coughlin, Roy, & Dusek, 2011). These observable measures can therefore be consulted to provide varying amounts of evidence of a distracted driver. Following this logic, it seems intuitive that an algorithm that learns from multiple vehicle and human variables that are individually sensitive to distraction would improve the sensitivity and specificity of distraction detection. However, few studies of driver distraction have explored machine learning algorithms that include more than one measure (e.g., vehicle and driver physiological measures) in a single algorithm. The studies that have included multiple measures (e.g., Liang & Lee, 2014; Liang, Lee, & Reyes, 2007; Liang, Reyes, & Lee, 2007) have focused on eye-gaze and vehicle control input measures.

Another factor to consider is that different sources of distraction should be mitigated with different solutions. Effective interventions that help a driver recover from distraction due to daydreaming or engaging in a purely cognitive task are not necessarily the best for combating distraction in sensorimotor tasks such as texting on a mobile device (Engström & Victor, 2009). It is therefore an additional challenge to develop "multiclass" machine learning algorithms that can not only detect whether a driver is distracted but provide some insight into the distraction source so that mitigations can be most appropriate for the context.

The current study takes advantage of a large multivariate set of driver behavioral and physiological data collected during human subjects driving simulation studies (Taamneh et al., 2017). Using these data, the performance of several advanced machine learning techniques is investigated toward two goals. The first goal is to develop an effective algorithm that (a) combines driver performance/vehicle input and physiological data sources; (b) is informed by domain knowledge in the generation and selection of complex features; and (c) is designed to distinguish among multiple classes of distraction (nondistraction, cognitive distraction, texting). To the authors' knowledge, such an algorithm has not been introduced in the literature. The second goal of this study is to mine successful predictive algorithms for new insights on driver distraction that can be used to inform mitigation technologies and future experiments. The remainder of this article reviews current detection algorithms, discusses the model training and evaluation process, and discusses the implications of the model findings.

Current Distraction Detection Algorithms

A substantial amount of research has been conducted on supervised machine learning algorithms for distraction detection (Ersal, Fuller, Tsimhoni, Stein, & Fathy, 2010; Li, Jain, & Busso, 2013; Liang & Lee, 2014; Liang, Lee, et al., 2007; Liang, Reyes, et al., 2007; Liu, Yang, Huang, Yeo, & Lin, 2016; Masood et al., 2018; Miyaji, Kawanaka, & Oguri, 2009; Sathyanarayana, Nageswaren, Ghasemzadeh, Jafari, & Hansen, 2008; Son & Park, 2016; Zhang, Owechko, & Zhang, 2004). The algorithms developed by this literature can be characterized by their input data, their ground truth definition of driver distraction, and machine learning approach. Input sources include driving behavior, head and eye tracking, and driver physiological measures. Prior algorithms have used one of two types of ground truth: binary (e.g., distracted cases and normal driving cases) and multiclass (e.g., cognitively distracted cases, texting cases, and normal driving cases). Algorithms that leverage multiple sources of input and multiple classes of ground truth have the most power for inference because supervised machine learning algorithms can only learn from the data and labels provided in their training data set. In this

TABLE 1: Summary of Multiclass Distraction Detection Algorithms

Study	Input Data	Machine Learning Approach	Feature Set	Ground Truth
Torkkola, Massey, and Wood (2004)	Steering angle Accelerator pedal input Lane metrics	Regression Tree	Mean Variance Entropy Stationarity	Multiclass gaze metrics
Li et al. (2013)	Head position Eye closure Speed Steering angle Brake pedal input	k-Nearest Neighbor SVM	Mean Standard deviation Maximum Minimum Range Interquartile range Skew Kurtosis Frequency Duration	Multiclass classifier with cognitive and visual secondary tasks
Son and Park (2016)	Lane position Steering angle	Neural Network	Standard deviation	Multiclass classifier with cognitive and visual secondary tasks
Masood et al. (2018)	Images of the driver and vehicle interior	Convolutional Neural Network	Spatial image features	Multiclass classifier with nine distracting behaviors

Note. SVM = Support Vector Machines.

context, prior work is limited as few approaches have investigated multiclass ground truth. Of the algorithms that have investigated multiclass ground truth, summarized in Table 1, the majority have focused on driver behavior input. The two exceptions, Li et al. (2013) and Masood et al. (2018), augmented driver behavior with head- and eye-tracking measures. Thus, there is a gap in the prior work with multiclass algorithms that use physiological or a combination of driver behavior and physiological measures.

Prior work has explored several machine learning approaches including Support Vector Machines (SVM; Ersal et al., 2010; Jin et al., 2012; Li et al., 2013; Liang, Reyes, et al., 2007; Liu et al., 2016; Miyaji et al., 2009), Bayesian Networks (Liang & Lee, 2014; Liang, Lee, et al., 2007), Neural Networks (NNs; Ersal et al., 2010; Masood et al., 2018; Son & Park, 2016), Decision Trees (DTs; Zhang et al., 2004), Random Forests (RFs; Ragab, Craye, Kamel, & Karray, 2014), and k-Nearest Neighbor (kNN) classifiers (Li et al., 2013; Sathyanarayana et al., 2008). No single approach has shown to significantly outperform all others, although there is some evidence that RFs may outperform DTs and SVM (Ragab et al., 2014). To the best of the authors' knowledge, there has not been a comprehensive comparison of these techniques. In addition, there has not been an attempt to use trained models to find new insights into distracted driving. This study addresses the gaps through a comprehensive comparison of algorithms and a variable importance analysis.

METHODS

Data Collection

The data set used in this analysis was collected in Texas A&M Transportation Institute's (TTI) Realtime Technologies Inc. (RTI) driving simulator (shown in Figure 1). The original goal

Figure 1. The experimental setup including simulator. Note the physiological data collection devices on the participant's wrist.

of the experiment was to evaluate physiological changes associated with types of distraction (Pavlidis et al., 2016). The experiment began with a total of 78 licensed, regular drivers from two age groups (18–27 years of age; older than 60 years of age); however, 1 participant withdrew due to motion sickness, data from 9 participants were lost due to technical issues, and physiological data were missing for 20 additional participants, resulting in 48 complete data sets. The data sets were approximately balanced across age and gender with 10 males and 14 females in the younger group, and 11 males and 13 females in the older group. The compiled data set from the experiment is published on the Open Science Framework (Taamneh et al., 2017).

Study process. The study included eight experimental phases separated by 2-min breaks, during which participants responded to questionnaires. The eight phases included a relaxation period with no driving task, a practice session on the simulator, a 15-min drive on a 10-km straight road (designed to relax the participants), four drives that included secondary task loading (loaded drives), and a 5-min final drive on a 3.2-km straight road that included a surprise unintended acceleration event. Figure 2 illustrates the full process. The secondary tasks in the loaded drives were characterized as loading *cognitive* (requiring a verbal response to mathematical or analytical questions), *emotional* (verbal response to emotionally charged questions), or *sensorimotor* (texting on a smartphone provided by the experimenters) resources. A fourth loaded drive included the same driving demands but no secondary task (the *normal* drive). The order of the four loaded drives was randomly counterbalanced across participants. The loaded drive portion of the study was designed as a 2 × 2 × 4 mixed design with Age group and Sex as between-subjects factors, and Load Type (normal, cognitive, emotional, sensorimotor) as a within-subjects factor. The analysis discussed here focused on the cognitive, sensorimotor, and normal loaded drives.

Simulator scenario. The driving scenarios in the four loaded drives were conducted on a 10.9-km section of a four-lane highway with a posted speed limit of 70 km/h, oncoming traffic density of 12 vehicles/km, and 2 buildings/km. Drivers were instructed to drive in the rightmost lane and periodically encountered construction zones in the adjacent lane. Approximately halfway through the drive (5.2 km), drivers were forced to change lanes due to road construction cones blocking the rightmost lane. Participants were instructed to perform secondary tasks (driving the vehicle remained the primary task throughout) during two nonconsecutive phases of the loaded drives. These phases began at 1.2 km into the drive and at 7.2 km into the drive and continued for 3.2 km (see Figure 3). Following the completion of the secondary task, drivers were instructed to resume normal driving. Each loaded drive took approximately 15 min to complete.

Data set. Driving behavior data and driver physiological measures were collected continuously throughout each drive. Driving behavior measures included instantaneous measures of acceleration, brake force, distance, lane offset, lane position, speed, and steering signals. The

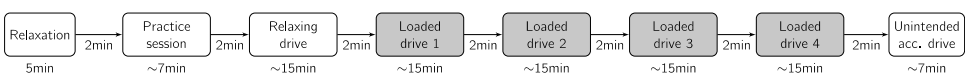


Figure 2. Temporal depiction of the eight experimental phases of the study. The drives included in this analysis are highlighted in gray. ACC = acceleration.

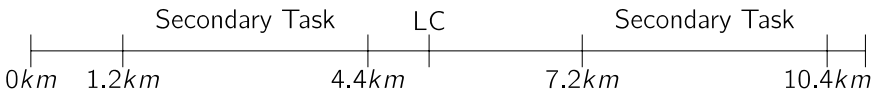


Figure 3. Drive segments by distance into the drive. The secondary task periods represent the portions of the drive where drivers were asked to engage in a secondary task (e.g., texting). LC = the point in the drive where drivers were instructed to perform a lane change.

TABLE 2: Sample of the Raw Data Set Used for Algorithm Training and Evaluations

Time	Secondary Task	Heart Rate	Breathing Rate	Perinasal Perspiration	Speed	Accelerator Pedal Angle	Brake	Steering	Lane Offset
75	0	99	16.2	0.0046	74	5.54	0	0.00	–0.54
76	0	101	16.2	0.0046	74	5.14	0	0.00	–0.50
77	0	99	16.0	0.0047	74	5.50	0	0.00	–0.45
78	0	95	16.0	0.0047	74	5.36	0	–0.01	–0.40
79	0	92	16.2	0.0048	74	4.44	0	–0.01	–0.36
80	1	89	16.2	0.0048	74	0.00	0	–0.01	–0.35
81	1	86	16.1	0.0048	73	0.00	0	0.00	–0.38

Note. Columns including session information (e.g., drive) are omitted for brevity.

driver physiological measures included perinasal perspiration (see Pavlidis et al., 2016), palm electrodermal activity, heart rate, breathing rate, and eye tracking data. The driving and physiological measures were collected at 60 Hz but downsampled to 1 Hz in the published data set. This analysis specifically focused on speed, lane offset, steering angle, brake pedal position, heart rate, perinasal perspiration, and breathing rate, because prior work has illustrated that these measures are sensitive to cognitive distraction, texting, or both. A sample of the full data set is shown in Table 2.

Data Preprocessing

The data set was processed in R (R core team, 2017) through four steps: normalization, windowing, window labeling, and separating into training and testing subsets. The normalization step consisted of subtracting the sample mean from each measure and dividing by the standard deviation. This step was necessary as some machine learning approaches explored here (e.g., SVM) are sensitive to unscaled data. The windowing step involved dividing each drive into nonoverlapping 30-s windows. The window labeling step consisted of assigning a label of cognitive distraction, texting, or normal driving to each 30-s window. Windows were labeled with either the majority class label or, in the case of a tie (which was the case in approximately 7% of the overall data set), the first label in the window was chosen. This method was chosen to account for delays in and lingering effects of the distracting secondary task. A sample of the normalized, windowed, and labeled data set is shown in Table 3. The full data set included 2,060 windows: 470 texting, 504 cognitively distracted, and 1,086 normal.

Finally, the data were randomly split into training and testing data sets, with the training set including 90% of the data and the testing set including 10%. The data were split by driver (rather than by window or by drive) to provide the best possible estimate of how the findings would generalize to an average driver. The testing set was approximately balanced across both age group and sex and contained 231 total windows—53 cognitively distracted, 129 normal driving, and 49 sensorimotor distraction. The training set was further downsampled to achieve an even class distribution. This step was

TABLE 3: Sample of the Normalized, Windowed, and Labeled Data

Window	Time	Train or Test	Label	Normalized HR	Normalized Breathing Rate	Normalized Speed	. . .
1	1	Train	Normal	0.81	0.39	0.38	. . .
1	2	Train	Normal	1.16	0.39	0.38	. . .
1	3	Train	Normal	0.81	0.36	0.38	. . .

Note. Additional columns including other measures are excluded for clarity. HR = heart rate.

TABLE 4: Example of the Reduced Feature Data Set

Window	Train or Test	Label	Lane Offset Standard deviation	Lane Offset Variance	Speed Maximum	. . .
1	Train	Normal	0.448	0.201	0.543	. . .
2	Train	Normal	0.465	0.216	0.035	. . .
3	Train	Normal	0.251	0.063	0.396	. . .

Note. Additional features are not included for clarity.

necessary to avoid bias in algorithm training, for example, algorithms that predominantly predict a single class (Kuhn & Johnson, 2013). After downsampling, the training set contained 421 windows of each class (1,263 total).

Feature Extraction and Reduction

Following data preprocessing, feature extraction and reduction were completed in Python using the Time Series Feature Extraction based on Scalable Hypothesis tests (TSFRESH) package (Christ, Kempa-Liehr, & Feindt, 2016). TSFRESH automatically extracts and filters hundreds of features, including distributional, nonlinear, spectral, Fourier, wavelet, polynomial, and other miscellaneous features such as the frequency of peaks. TSFRESH first calculates all possible features and then performs feature filtering based on a multiple test procedure from the theory of hypothesis testing, which is explained in detail in Christ et al. (2016) and in Benjamini and Yekutieli (2001). The feature reduction step removes both exceptionally rare and exceptionally common feature values from the data. This step reduces the likelihood of identifying coincidental events, such as an animal crossing the road in front of a driver, as distraction events, and improves the speed of algorithm training. After obtaining the output data sets from TSFRESH, additional feature reduction was performed in R. The additional feature reduction included removing features with missing or infinite values, features with zero or near zero variance, and features with high correlations to other features, resulting in a total of 438 features included in each window of the training and testing data sets. A sample of the reduced data set is shown in Table 4.

Algorithm Training and Evaluation

This analysis sought to determine the relative impacts of driver behavioral and physiological features, and of machine learning approaches, on model prediction performance. To accomplish this goal, we analyzed and compared three different data sets for algorithm input:

1. Physiological Data Set: including only driver physiological measures (e.g., breathing rate, heart rate, and perinasal perspiration).
2. Driver Behavior Data Set: including only driving behavior measures (e.g., brake force, lane offset, speed, and steering angle).
3. Combined Data Set: including both driving behavior and driver physiological measures.

In addition to the input sets, seven different machine learning approaches—selected based

on the techniques applied in prior literature—were used. The approaches and associated implementation packages are as follows:

1. RF (Liaw & Weiner, 2002)
2. DT (Therneau, Atkinson, & Ripley, 2017)
3. Naïve Bayes (NB; Majka, 2018)
4. kNN (Schliep & Hechenbichler, 2016)
5. SVM with linear kernel function (svmLin; Meyer et al., 2017)
6. SVM with a radial kernel function (svmRad; Karatzoglou, Smola, Hornik, & Zeileis, 2004)
7. NN (Venables & Ripley, 2002)

In total, 21 algorithms were developed, one for each combination of input data type and machine learning approach. All algorithms were trained using the functions in the caret package in R (Kuhn et al., 2017).

RESULTS

The algorithms were compared across their accuracy, average binary class area under the receiver operating characteristic (ROC) curve (Fawcett, 2004), and their confusion matrices for their predictions on the testing data set. Following the algorithm performance analysis, variable importance calculations for the top performing algorithm were used to provide additional insights into behavioral patterns during distracted driving.

Algorithm Classification Performance

The seven machine learning approaches were trained using each of the three input data types and assessed with the testing data set. Figure 4 shows the test set accuracy, along with bootstrapped 95% confidence intervals, and Figure 5 shows the mean test set area under the curve (AUC) for each algorithm, grouped by input data type. In both figures, random guessing performance is shown as a horizontal black line. The SVM algorithm implemented with a linear kernel (i.e., svmLin) is not included in the AUC values because the R implementation provides categorical output and thus thresholding is not possible. The algorithm accuracy differences were statistically evaluated using McNemar's test (Dietterich, 1998) with a threshold of $p < .05$.

Figure 4 illustrates that all but 6 of the 21 algorithms were significantly more accurate

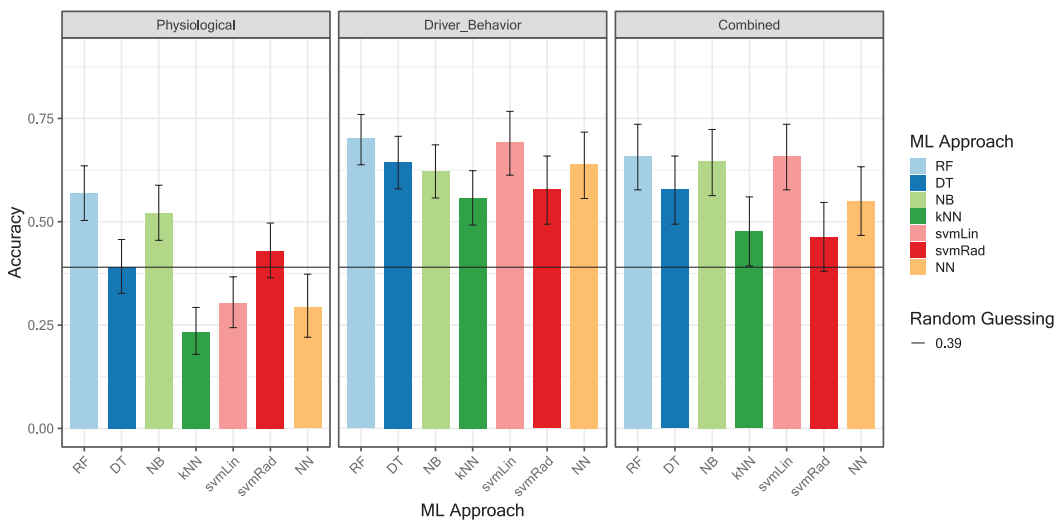


Figure 4. Algorithm accuracy arranged by input measures and ML approach. The black line indicates the accuracy of a random classifier. The error bars represent 95% confidence intervals. ML = machine learning; RF = Random Forest; DT =Decision Tree; NB = Naïve Bayes; kNN = k-Nearest Neighbor; svmLin = Support Vector Machine with linear kernel function; svmRad = Support Vector Machine with a radial kernel function; NN = Neural Network.

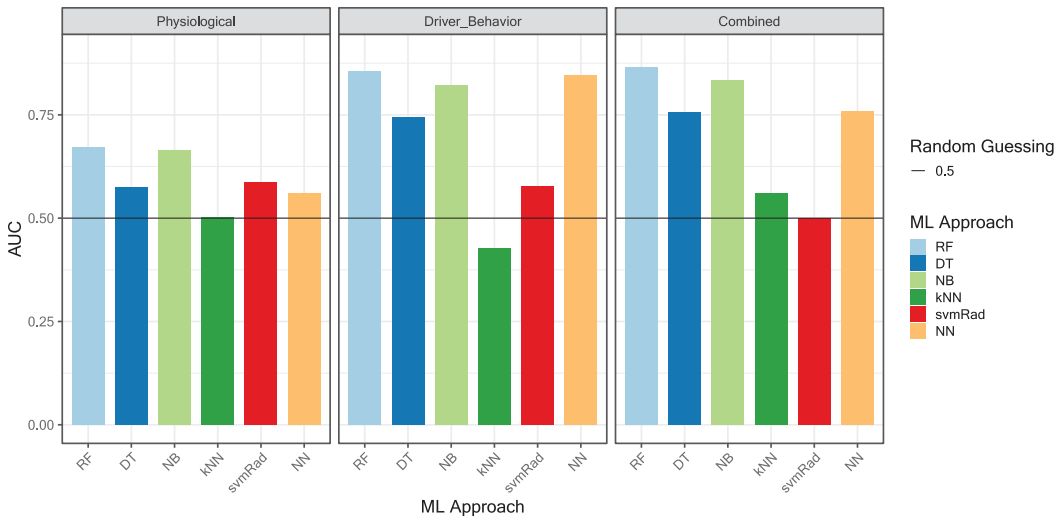


Figure 5. Algorithm average binary AUC arranged by input measures and ML approach. The black line indicates the AUC of a random classifier. Note that confidence intervals are not provided due to the calculation method (Hand & Till, 2001). AUC: area under the ROC curve; ML = machine learning; RF = Random Forest; DT =Decision Tree; NB = Naïve Bayes; kNN = k-Nearest Neighbor; svmRad = Support Vector Machine with a radial kernel function; NN = Neural Network.

than random guessing (all $p < .001$). The 6 algorithms comparable with random guessing used driver physiological data or a combination of physiological data and driver behavior data as input (Physiological DT, kNN, SVM Linear, SVM Radial, and NN, Combined SVM Radial). The RF and SVM with a linear kernel using driver behavior input had the highest accuracy. Pairwise comparisons across the RF models showed that algorithms including driver behavior measures significantly outperformed the physiological algorithm (all $p < .001$). Furthermore, there was no significant difference in accuracy between only driver behavior–based algorithms and the combined algorithms. This result suggests that driving behavior measures dominate physiological measures in the task of classifying driver distraction.

The results in Figures 4 and 5 indicate that some notion of vehicle control is necessary for identifying distraction. More interestingly, the high AUC values in Figure 5 suggest that the driving behavior measures used in this study (i.e., brake force, lane offset, speed, and steering angle) are sufficient to differentiate between external, physical types of distraction such as texting, and internal, mental types of distraction such as solving analytical problems (*cognitive* distraction). A final observation in these figures is that although there were no significant differences between machine learning algorithms, the RF method tended to provide more consistent results across all input types and produced the most accurate algorithm. For this reason and for brevity, the remaining analyses will focus on the RF-based algorithms.

Further insight into the accuracy and AUC results can be gained through analysis of the complete prediction set depicted as a confusion matrix. Figure 6 shows the confusion matrix for the RF algorithms built on Physiological, Driver Behavior, and Combined input data sets. The figure shows the proportion of total training instances (i.e., windows) that were classified correctly (on the diagonal, highlighted in gray) and incorrectly by the algorithms for each class. One unique finding from the figure is that although the physiological algorithm had similar accuracy to random guessing for detecting texting (shown in the bottom right chart of Figure 6), it had comparable accuracy to the other algorithms in detecting cognitive distraction (shown

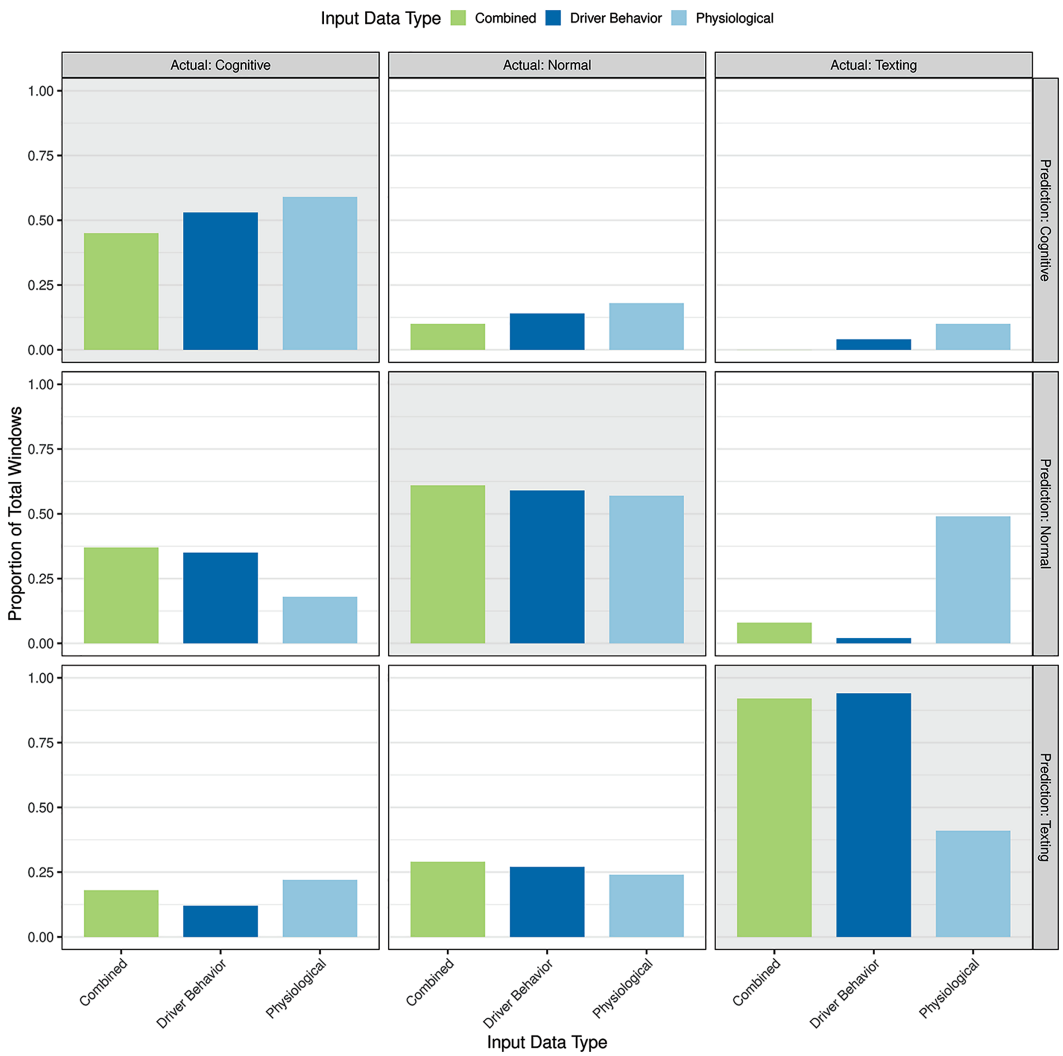


Figure 6. Confusion matrix for Random Forest algorithms and input measures. The gray highlighted charts on the diagonal indicate correct predictions and the off-diagonal plots indicated incorrect predictions.

in the top left chart). In the cases where drivers were actually texting and the algorithms predicted incorrectly (shown in the first and second rows, far right column), the physiological algorithm most often confused the texting cases with normal driving.

Inference With Variable Importance

Although the algorithm performance metrics suggest that driver behavior algorithms can differentiate distraction, they do not support analysis of the relative impacts of features. Variable importance measures, which estimate the mean decrease in accuracy associated with removing a feature from the algorithm, are one way of addressing this gap. Figure 7 shows the 10 most important variables for the RF driver behavior algorithm and Table 5 provides an explanation of each feature.

The results show that lane offset, speed, and steering are the most important measures—that is, they are the most sensitive measures to distraction. Standard deviation of lane offset is the most important feature by a substantial margin, although some nuanced measures including the number of speed measurements greater than the

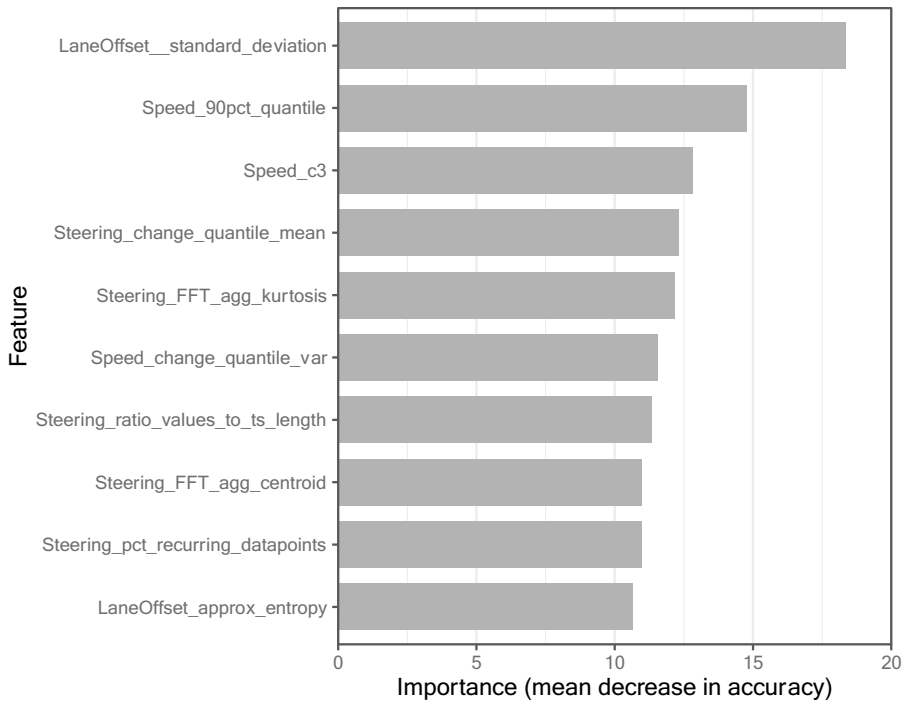


Figure 7. Variable importance values—measured by mean decrease in accuracy associated with removing the feature—for the 10 most important features in the Random Forest driver behavior algorithm. FFT = Fast Fourier Transform.

90th percentile quantile, the C3 measure of speed, and the change quantiles of steering also have substantial impact on algorithm performance. These features are, for the most part, not represented in prior analyses of distraction behavior including distraction algorithm development.

Beyond these findings, further insight can be achieved by analyzing the distribution of features across classes. Although analyzing these distributions in isolation negates the power of machine learning methods to find complex patterns across features, it provides some insight to the patterns in the data associated with classes. Figure 8 shows an example of this analysis—a violin plot—partitioned by feature and distraction type and ordered by variable importance. Each violin plot shows the distribution a feature mirrored across the horizontal axis (Hintze & Nelson, 1998). Differences in the means (shown as points on the plot) or the shapes of the distributions between the classes suggest that the measures may be useful indicators of distraction.

The plots show clear differences in the distributions of data associated with cognitive distraction, normal driving, and texting in the four most important variables (i.e., the top four charts in the figure). In particular, the standard deviation of lane offset chart shows that texting drivers had an almost uniform distribution, whereas cognitively distracted and normal drivers were more normally distributed. The mean standard deviation of lane offset is lowest in the cognitively distracted cases and highest in the texting cases. The 90th percentile speed plot shows a lower mean value for texting compared with normal driving and cognitive distraction and a broader variance. Thus, texting drivers, on average, have fewer instances of speeds greater than the 90th percentile than normal drivers. In both cases, the differences in the distributions suggest that these metrics would be viable indicators of distraction, even for analyses that do not use

TABLE 5: Description of the Most Important Features of the Random Forest Driver Behavior Algorithm

Measure	Feature Type	Description
Lane offset	Standard deviation	Standard deviation of the time series.
Speed	90% quantile	The number of data points greater than the 90% quantile of the time series.
Speed	C3	A measure of nonlinearity of a time series calculated as the autocovariance of the current and two previous values of the time series (Schreiber & Schmitz, 1997).
Steering	Change quantile mean	The average absolute value of the changes in the time series.
Steering	FFT aggregated kurtosis	The kurtosis of the FFT of the time series.
Speed	Change quantile variance	The variance of the absolute value of the consecutive changes in the time series.
Steering	Ratio value to time-series length	The ratio of the number of unique values to the total length of the time series.
Steering	FFT aggregated centroid	The mean frequency of the FFT of the time series.
Steering	Percentage of recurring data points	The percentage of unique values in the time series. The value is 1 if all values are unique and less than 1 if multiple values are repeated.
Lane offset	Approximate entropy	A measure of the level of randomness of the time series.

Note. FFT = Fast Fourier Transform.

machine learning, such as analysis of variance (ANOVA). As the variable importance decreases (e.g., in the lane offset approximate entropy variable), there is more overlap in the distributions and thus it is less likely that these variables would show significant differences in empirical analyses using linear models. However, pairwise Kolmogorov–Smirnov tests—which assess whether samples of data originate from different distributions—showed that all of the distributions shown in Figure 8 are significantly different (all $p < .001$). Collectively, these results indicate that there may be a benefit to using 90th percentile quantiles, nonlinearity (C3), and change quantile metrics as dependent measures of distraction in subsequent analyses. Such measures may be more sensitive to distraction than other more common metrics such as mean speed and steering reversal rates, which were also included in this analysis but found to be less important for classification.

DISCUSSION

Distracted driving remains one of the largest safety risks to current ground transportation systems (Dingus et al., 2016; NCSA, 2017). Although past work has advanced the power of distraction detection algorithms, few attempt multiclass classification (e.g., Li et al., 2013), and fewer systematically compare advanced machine learning techniques. The classification aspect is important for effectively mitigating distraction, as cognitive distraction and texting-related distraction should be mitigated with different strategies. Very few studies published in driving contexts have compared and contrasted input data sets that included physiological, behavioral, or both types of data, and to the authors' knowledge, none have attempted this comparison with the inclusion of complex, nonlinear features that may be maximally sensitive to certain types of distraction. The results of this work can inform human performance modeling and driver distraction algorithm development and illustrate the value of advanced machine learning techniques for inferring human states.

Algorithm Classification Performance

The performance of distraction classification algorithms was affected more by the selected

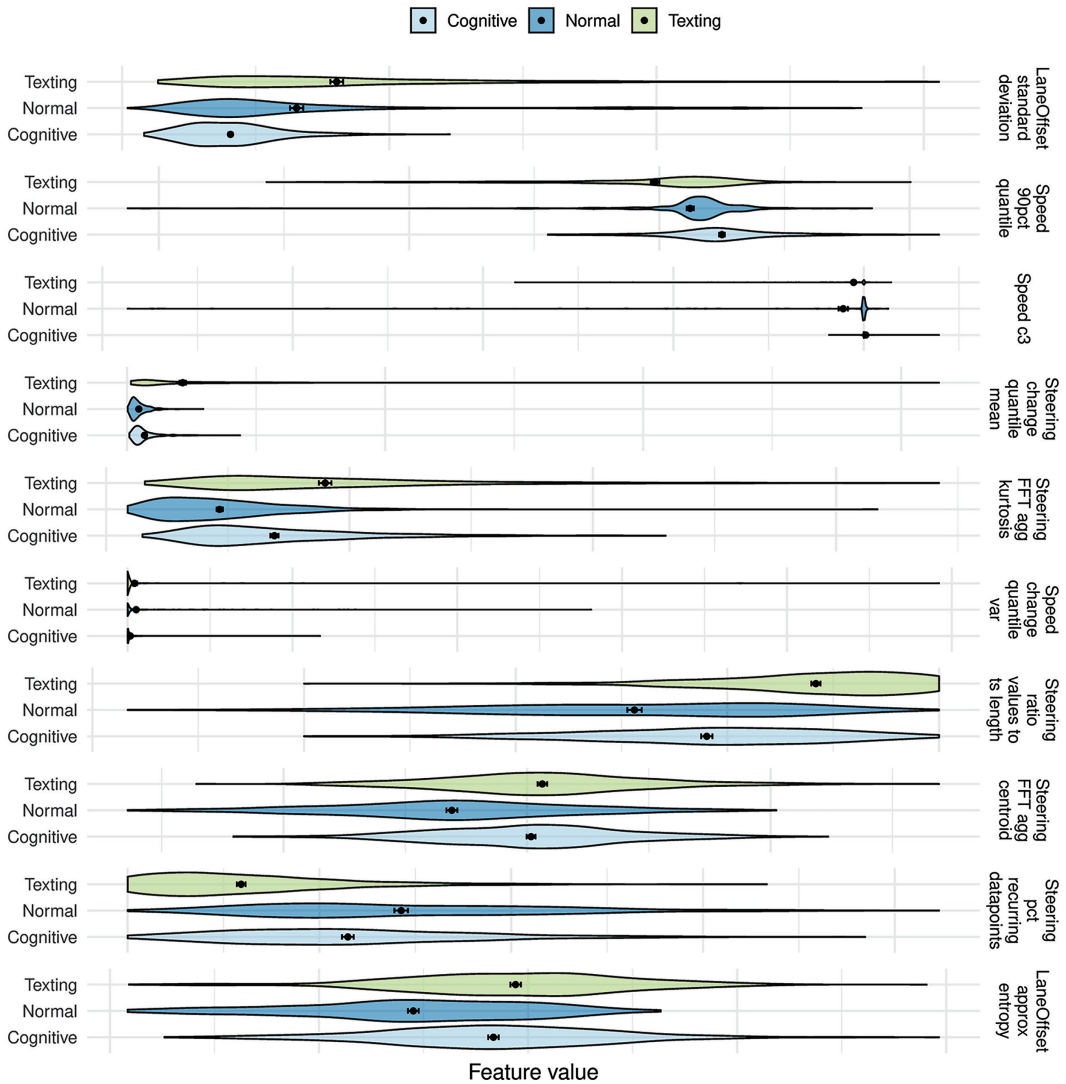


Figure 8. Violin plots of feature values by distraction condition. FFT = Fast Fourier Transform.

input measures (e.g., driving performance and/or physiological data, as well as generated features) than by the machine learning approaches explored here. The accuracy and AUC results show that driver behavior (or a combination of driver behavior and physiological) measures are effective for differentiating cognitive distraction and texting from normal driving. Algorithms that use physiological measures alone were able to differentiate normal driving from states of cognitive distraction, but not from sensorimotor distraction (texting). This is a noteworthy limitation, as distracting activities that engage sensorimotor resources are associated with some of the largest increases in crash risk (Caird, Johnston, Willness, Asbridge, & Steel, 2014; Caird, Simmons, Wiley, Johnston, & Horrey, 2018; Klauer et al., 2014) and thus should be emphasized in the development of distraction detection and mitigation systems.

The findings from the current study add to a mixed body of evidence from research seeking to infer operator cognitive states by combining and comparing various performance-based, subjective, and physiological measures. In a driving simulator, Hicks and Wierwille (1979) compared

subjective measures, driver behavioral data, and physiological correlates of workload. Similar to the results of the current study, driving measures—specifically, steering and lateral deviation metrics—were found to be considerably more sensitive than physiological measures to the effects of imposed workload. Similarly, Engström, Johansson, and Östlund (2005) and Pavlidis et al. (2016) showed that when task-imposed workload is increased in a driving simulation, lateral and longitudinal control measures are reliably sensitive to changes in workload, whereas physiological measures are comparably less sensitive.

In contrast, other research has found physiological measures to be highly sensitive to changes in stress or imposed task load on drivers (Brookhuis & de Waard, 2010). Healey and Picard (2005) showed that sophisticated algorithms based on physiological measures can distinguish up to three levels of driver stress with accuracies as high as 97%. Mehler, Reimer, and Coughlin (2012) found greater sensitivities for physiological measures than for driving-performance measures when discriminating among levels of cognitive demand. Similarly, Solovey, Zec, Perez, Reimer, and Mehler (2014) found an improvement in workload discrimination when physiological data were included in machine learning algorithms along with driver performance data.

The inconsistency in historical results comparing physiological and performance-based indicators of driver cognitive state may be partially due to methodological factors such as participant selection, training methods, task selection and pacing, and the scope of allowable driver behaviors (Mehler et al., 2012). The inconsistency may also reflect the relationship between imposed task loads and one's personal cognitive load limit, sometimes referred to as the "redline" of cognitive workload (Grier et al., 2008; Wickens, Hollands, Banbury, & Parasuraman, 2013). Although task loads are below one's redline, performance is often only minimally affected by changing load levels, especially when tasks are characterized by high levels of automaticity (Engström et al., 2017), because drivers have resources available to engage as more support is needed (Mehler et al., 2012). At these "sub-redline" workload levels, driving-performance measures are therefore less sensitive than physiological indicators of workload. This logic can be applied to the development of physiological algorithms that can detect and trigger mitigation strategies when operator workloads approach the redline (Rodriguez-Paras, Susindar, Lee, & Ferris, 2017; Rodriguez-Paras, Yang, Tippey, & Ferris, 2015) to minimize performance and safety decrements. Indeed, the accuracy and AUC results from this study suggest that algorithms that combine driver behavior and physiological changes may be robust to this redline performance; however, the results should be confirmed by a more rigorous analysis.

Inferential Analysis

The variable importance analysis found that lane offset, speed, and steering provide the most sensitive measures in the distraction classification algorithm. Lane offset standard deviation is the most sensitive metric, followed by the frequency of speeds greater than the 90% quantile, the nonlinearity of the speed (C3), the mean of the absolute value of changes in steering (change entropy), and the aggregated kurtosis of the Fast Fourier Transform (FFT) of the steering signal. These findings align with prior work that has indicated that both cognitive and texting distractions affected lane position variance (Cooper, Medeiros-Ward, & Strayer, 2013; Drews et al., 2009; Engström et al., 2005, 2017; Liang & Lee, 2010). Although the current study included simple steering metrics commonly used to assess distraction, such as steering reversal rate (Macdonald & Hoffmann, 1980) or steering entropy (Boer, 2000), the feature importance analysis showed that more complex, nonlinear transformations of steering data, such as steering change quantiles and FFT metrics, were more sensitive to the types of distraction investigated in this work.

These findings are noteworthy because they indicate how nonlinear complex features may offer a clear direction for improving on existing driver distraction algorithms and analyses built on simpler metrics. For example, there has been considerable divergence in how measures of speed (as well as other performance measures) are associated with different types of distraction

(e.g., Engström et al., 2005, 2017). The findings here offer the possibility that some of this divergence is related to the manner in which speed is represented in models. Although most prior studies have examined linear patterns in mean speed, the current study found speed measures above the 90% quantile to be far more sensitive and specific for classifying states of distraction. Further research can investigate whether such advanced metrics may be more robust to individual differences in driving styles, which represent a key challenge in this research (Engström et al., 2005).

The algorithms in the current work were designed to distinguish three driver states which included normal driving and driving while distracted by either cognitive or sensorimotor-based secondary tasks. Most prior work focuses on unidimensional assessments that are primarily quantitative, designed to support additive comparisons in workload levels (e.g., "high" vs. "low" levels). In contrast, the current classification effort attempts to qualitatively infer the nature of the distraction by comparing measures that respond differently when loading cognitive, physical, and visual resources (e.g., Brookhuis & de Waard, 2010; Engström et al., 2005). Future research may explore a wider set of distracting tasks with special attention given to machine learning input features that are differentially sensitive to different types of loading.

Limitations and Future Work

The limitations of this analysis include the simulator and study design, the instrumentation and sensors used, and the machine learning approaches explored. The data were collected in realistic, but simulator-based scenarios. Although many prior studies support the validity of simulators for assessing behavioral and physiological changes associated with driver distraction (e.g., Eriksson, Banks, & Stanton, 2017; Mullen, Charlton, Devlin, & Bedard, 2011), several have observed differences in physiological and driver behavior measures between simulators and on road environments (e.g., Engström et al., 2005). Thus, one should be cautious about generalizing these findings beyond the task or environment explored here. Another limitation lies in the quality of data received from the sensors. Most of the data exclusions in this analysis were due to failures in the physiological sensor hardware or software. Although this in some ways strengthens the argument against a broad reliance on such sensors in distraction detection, it also limits the conclusions which can be made from these results. Finally, despite taking multiple precautions, there is some risk of model overfitting given the size of the training data and the number of parameters explored.

These limitations may be addressed through subsequent naturalistic driving studies that include a broader set of distractions, more-varied driving scenarios, and more opportunities to reliably collect accurate physiological data. Future work should also explore additional optimization techniques and more recent machine learning approaches, such as deep learning neural networks.

Application

The findings can be applied to both future technology development and empirical study design. Development of distraction mitigation systems should leverage these results to focus on driver behavior–based algorithms, in particular emphasizing lane position, speed, and steering behavior as input, as well as encouraging SVM or RF approaches and the use of diverse feature sets. Future empirical studies should consider the inclusion of lane offset standard deviation as well as quantile-based and nonlinear steering and speed metrics as dependent measures of distraction.

CONCLUSION

This study developed and tested algorithms for classifying texting, cognitive distraction, and normal driving using a mix of driver behavior and physiological input measures, a comprehensive set of feature types, and several machine learning approaches. The findings suggest that driver behavior metrics, combined with RF or SVM methods, and a diverse feature set are most promising for classifying distraction.

KEY POINTS

- Driver behavior and physiological measures may be used to create distraction classification algorithms.

- Overall, algorithms using driver behavior significantly outperform algorithms based solely on physiological measures.
- Physiological algorithms may be effective for identifying cognitive distraction but are not effective for detecting texting.
- Standard deviation of lane offset is the most sensitive feature to distraction along with speed and steering quantile measures.

ORCID iDs

Anthony D. McDonald https://orcid.org/0000-0001-7827-8828

Tyler A. Wiener https://orcid.org/0000-0003-2094-9638

REFERENCES

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165–1188. doi:10.1214/aos/1013699998

Boer, E. R. (2000). Behavioral entropy as an index of workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*, 125–128. doi:10.1177/154193120004401702

Brookhuis, K. A., & de Waard, D. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis and Prevention*, *42*, 898–903. doi:10.1016/j.aap.2009.06.001

Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014). A meta-analysis of the effects of texting on driving. *Accident Analysis & Prevention*, *71*, 311–318. doi:10.1016/j.aap.2014.06.005

Caird, J. K., Simmons, S. M., Wiley, K., Johnston, K. A., & Horrey, W. J. (2018). Does talking on a cell phone, with a passenger, or dialing affect driving performance? An updated systematic review and meta-analysis of experimental studies. *Human Factors*, *60*, 101–133. doi:10.1177/0018720817748145

Christ, M., Kempa-Liehr, A. W., & Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. Retrieved from http://arxiv.org/abs/1610.07717

Collet, C., Guillot, A., & Petit, C. (2010). Phoning while driving I: A review of epidemiological, psychological, behavioural and physiological studies. *Ergonomics*, *53*, 589–601. doi:10.1080/00140131003672023

Cooper, J. M., Medeiros-Ward, N., & Strayer, D. L. (2013). The impact of eye movements and cognitive workload on lateral position variability in driving. *Human Factors*, *55*, 1001–1014. doi:10.1177/0018720813480177

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1923. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9744903

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113, 2636–2641. doi:10.1073/pnas.1513271113

Dong, Y., Hu, Z., Uchimura, K., & Murayama, N. (2011). Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, *12*, 596–614. doi:10.1109/TITS.2010.2092770

Drews, F. A., Yazdani, H., Godfrey, C. N., Cooper, J. M., & Strayer, D. L. (2009). Text messaging during simulated driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*, 762–770. doi:10.1177/0018720809353319

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*, 97–120. doi:10.1016/j.trf.2005.04.012

Engström, J., Markkula, G., Victor, T., & Merat, N. (2017). Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human Factors*, *59*, 734–764. doi:10.1177/0018720817690639

Engström, J., & Victor, T. W. (2009). Real-time distraction countermeasures. In M. Regan, J. Lee, & K. L. Young (Eds.), *Driver distraction: Theory effects and mitigation* (pp. 465–484). Boca Raton, FL: CRC Press.

Eriksson, A., Banks, V. A., & Stanton, N. A. (2017). Transition to manual: Comparing simulator with on-road control transitions. *Accident Analysis & Prevention*, *102*, 227–234. doi:10.1016/j.aap.2017.03.011

Ersal, T., Fuller, H. J. A., Tsimhoni, O., Stein, J. L., & Fathy, H. K. (2010). Model-based analysis and classification of driver distraction under secondary tasks. *IEEE Transactions on Intelligent Transportation Systems*, *11*, 692–701. doi:10.1109/TITS.2010.2049741

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, *31*, 1–38.

Feng, F., Bao, S., Sayer, J. R., Flannagan, C., Manser, M., & Wunderlich, R. (2017). Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. *Accident Analysis & Prevention*, *104*, 125–136. doi:10.1016/j.aap.2017.04.012

Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., & St John, M. (2008). The red-line of workload: Theory, research, and design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*, 1204–1208. doi:10.1177/154193120805201811

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, *45*, 171–186. doi:10.1023/A:1010920819831

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*, 156–166. doi:10.1109/TITS.2005.848368

Hicks, T. G., & Wierwille, W. W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *21*, 129–143. doi:10.1177/001872087902100201

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*, 181–184. doi:10.1080/00031305.1998.10480559

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*, 196–205. doi:10.1518/001872006776412135

Jin, L., Niu, Q., Hou, H., Xian, H., Wang, Y., & Shi, D. (2012). Driver cognitive distraction detection using driving performance measures. *Discrete Dynamics in Nature and Society, 2012*, Article 432634. doi:10.1155/2012/432634

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab—An S4 package for kernel methods in R. *Journal of Statistical Software*, *11*. doi:10.18637/jss.v011.i09

Kim, S., Chun, J., & Dey, A. K. (2015). Sensors know when to interrupt you in the car. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems—CHI '15* (pp. 487–496). doi:10.1145/2702123.2702409

Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *The Journal of Emergency Medicine*, *46*, 600–601. doi:10.1016/j.jemermed.2014.02.017

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268. doi:10.1115/1.1559160

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (1st ed.). New York, NY: Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Hunt, T. (2017). *Package "caret": Classification and regression training*. Retrieved from https://cran.r-project.org/web/packages/caret/caret.pdf

Lee, J. D., Moeckli, J., Brown, T. L., Roberts, S. C., Schwarz, C. W., Yekhshatyan, L., . . . Davis, C. (2013). *Distraction detection and mitigation through driver feedback appendices* (Report No. DOT HS 811 547A). Washington, DC: National Highway Traffic Safety Administration.

Li, N., Jain, J. J., & Busso, C. (2013). Modeling of driver behavior in real world scenarios using multiple noninvasive sensors. *IEEE Transactions on Multimedia*, *15*, 1213–1225. doi:10.1109/TMM.2013.2241416

Liang, Y., & Lee, J. D. (2010). Combining cognitive and visual distraction: Less than the sum of its parts. *Accident Analysis & Prevention*, *42*, 881–890. doi:10.1016/j.aap.2009.05.001

Liang, Y., & Lee, J. D. (2014). A hybrid Bayesian network approach to detect driver cognitive distraction. *Transportation Research Part C: Emerging Technologies*, *38*, 146–155. doi:10.1016/j.trc.2013.10.004

Liang, Y., Lee, J. D., & Reyes, M. L. (2007). Nonintrusive detection of driver cognitive distraction in real time using Bayesian networks. *Transportation Research Record: Journal of the Transportation Research Board*, *2018*, 1–8. doi:10.3141/2018-01

Liang, Y., Reyes, M. L., & Lee, J. D. (2007). Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems*, *8*, 340–350. doi:10.1109/TITS.2007.895298

Liaw, A., & Weiner, M. (2002). Classification and regression by random forest. *R News*, 18–22. Retrieved from http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Liu, T., Yang, Y., Huang, G., Yeo, Y. K., & Lin, Z. (2016). Driver distraction detection using semi-supervised machine learning. *IEEE Transactions on Intelligent Transportation Systems*, *17*, 1108–1120. doi:10.1109/TITS.2015.2496157

Macdonald, W. A., & Hoffmann, E. R. (1980). Review of relationships between steering wheel reversal rate and driving task demand. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *22*, 733–739. doi:10.1177/001872088002200609

Majka, M. (2018). *naivebayes: High performance implementation of the naive bayes algorithm*. Retrieved from https://cran.r-project.org/package=naivebayes

Masood, S., Rai, A., Aggarwal, A., Doja, M. N., & Ahmad, M. (2018). Detecting distraction of drivers using convolutional neural network. *Pattern Recognition Letters*. Advance online publication. doi:10.1016/j.patrec.2017.12.023

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*, 6–12. doi:10.3141/2138-02

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*, 396–412. doi:10.1177/0018720812442086

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2017). *R Package e1071 Version 1.6-8. Gpl-2*. Retrieved from https://cran.r-project.org/web/packages/e1071/e1071.pdf

Miyaji, M., Kawanaka, H., & Oguri, K. (2009). Driver's cognitive distraction detection using physiological features by the AdaBoost. In *IEEE conference on intelligent transportation systems (ITSC)* (pp. 90–95). doi:10.1109/ITSC.2009.5309881

Mullen, N., Charlton, J., Devlin, A., & Bedard, M. (2011). Simulator validity: Behaviors observed on the simulator and on the road. In D. L. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of driving simulation for engineering, medicine, and psychology* (pp. 1–18). Boca Raton, FL: CRC Press.

National Center for Statistics and Analysis. (2017, March). *Distracted driving 2015*. (Traffic Safety Facts Research Note. Report No. DOT HS 812 381). Washington, DC: National Highway Traffic Safety Administration.

Pavlidis, I., Dcosta, M., Taamneh, S., Manser, M., Ferris, T., Wunderlich, R., & Tsiamyrtzis, P. (2016). Dissecting driver behaviors under cognitive, emotional, sensorimotor, and mixed stressors. *Scientific Reports*, *6*, 1–12. doi:10.1038/srep25651

Ragab, A., Craye, C., Kamel, M. S., & Karray, F. (2014, October). *A visual-based driver distraction recognition and detection using random forest*. Paper presented at the International Conference Image Analysis and Recognition, Vilamoura, Portugal (pp. 256–265).

R core team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org

Reimer, B., Mehler, B., Coughlin, J. F., Roy, N., & Dusek, J. A. (2011). The impact of a naturalistic hands-free cellular phone task on heart rate and simulated driving performance in two age groups. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*, 13–25. doi:10.1016/j.trf.2010.09.002

Rodriguez-Paras, C., Susindar, S., Lee, S., & Ferris, T. K. (2017). Age effects on drivers' physiological response to workload. *Proceedings of the Human Factors and Ergonomics Society, 61*, 1886. doi:10.1177/1541931213601951

Rodriguez-Paras, C., Yang, S., Tippey, K., & Ferris, T. K. (2015). Physiological indicators of the cognitive redline. *Proceedings of the Human Factors and Ergonomics Society*, *59*, 637–641. doi:10.1177/1541931215591139

Sathyanarayana, A., Nageswaren, S., Ghasemzadeh, H., Jafari, R., & Hansen, J. H. L. (2008). Body sensor networks for driver distraction identification. In *2008 IEEE international conference on vehicular electronics and safety* (pp. 120–125). New York, NY: IEEE. doi:10.1109/ICVES.2008.4640876

Schliep, K., & Hechenbichler, K. (2016). *kknn: Weighted k-nearest neighbors*. Retrieved from https://cran.r-project.org/package=kknn

Schreiber, T., & Schmitz, A. (1997). Discrimination power of measures for nonlinearity in a time series. *Physical Review E*, *55*, 5443–5447. doi:10.1103/PhysRevE.55.5443

Schwarz, C., Brown, T., Lee, J., Gaspar, J., & Kang, J. (2016). The detection of visual distraction using vehicle and driver-based sensors. In *SAE 2016 world congress and exhibition*, Detroit, MI. doi:10.4271/2016-01-0114

Smith, M. R. H., Witt, G. J., Bakowski, D. L., Leblanc, D., & Lee, J. D. (2009). Adapting collision warnings to real-time estimates of driver distraction. In M. Regan, J. D. Lee, & K. L. Young (Eds.), *Driver distraction: Theory effects and mitigation* (pp. 501–518). Boca Raton, FL: CRC Press.

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems—CHI '14* (pp. 4057–4066). New York, NY: ACM Press. doi:10.1145/2556288.2557068

Son, J., & Park, M. (2016, May). *Real-time detection and classification of driver distraction using lateral control performance*. Paper presented at the ACCSE 2016: The First International Conference on Advances in Computation, Communications and Services, Valencia, Spain.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, *9*, 23–32. doi:10.1037/1076-898X.9.1.23

Taamneh, S., Tsiamyrtzis, P., Dcosta, M., Buddharaju, P., Khatri, A., Manser, M., . . . Pavlidis, I. (2017). A multimodal dataset for various forms of distracted driving. *Scientific Data*, *4*, Article 170110. doi:10.1038/sdata.2017.110

Therneau, T., Atkinson, B., & Ripley, B. (2017). *rpart: Recursive partitioning and regression trees*. Retrieved from https://cran.r-project.org/package=rpart

Tippey, K. G., Sivaraj, E., & Ferris, T. K. (2017). Driving while interacting with Google glass: investigating the combined effect of head-up display and hands-free input on driving safety and multitask performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*, 671–688. doi:10.1177/0018720817691406

Torkkola, K., Massey, N., & Wood, C. (2004). Detecting driver inattention in the absence of driver monitoring sensors. In *Proceedings of the 2004 International Conference on Machine Learning and Applications* (pp. 220–226). New York, NY: IEEE. doi:10.1109/ICMLA.2004.1383517

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (4th ed.). New York, NY: Routledge.

Zhang, Y., Owechko, Y., & Zhang, J. (2004). Driver cognitive workload estimation: a data-driven perspective. In *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems* (IEEE Cat. No. 04TH8749, pp. 642–647). New York, NY: IEEE. doi:10.1109/ITSC.2004.1398976

Anthony D. McDonald is an assistant professor of industrial and systems engineering at Texas A&M University and directs the Human Factors and Machine Learning Laboratory. He received his PhD in industrial engineering from the University of Wisconsin–Madison in 2014.

Thomas K. Ferris is an associate professor of industrial and systems engineering at Texas A&M University, where he is the director of the Human Factors and Cognitive Systems Laboratory. He received his PhD in industrial and operations engineering from the University of Michigan in 2010.

Tyler A. Wiener is a supply chain engineer at Walmart Incorporated. He earned his BS in industrial engineering from Texas A&M University in 2018. The present paper was authored while he was an undergraduate researcher in the Human Factors and Machine Learning Laboratory at Texas A&M University.

Date received: July 23, 2018
Date accepted: May 15, 2019

JUST PUBLISHED!

Usability Assessment: How to Measure the Usability of Products, Services, and Systems

Volume 1, Users' Guides to Human Factors and Ergonomics Methods

By Philip Kortum

Usability Assessment is a concise volume for anyone requiring knowledge and practice in assessing the usability of any type of product, tool, or system *before* it is launched. It provides a brief history and rationale for conducting usability assessments and examples of how usability assessment methods have been applied, takes readers step by step through the process, highlights challenges and special cases, and offers real-life examples. By the end of the book, readers will have the knowledge and skills they need to conduct their own usability assessments without requiring that they read textbooks or attend workshops.

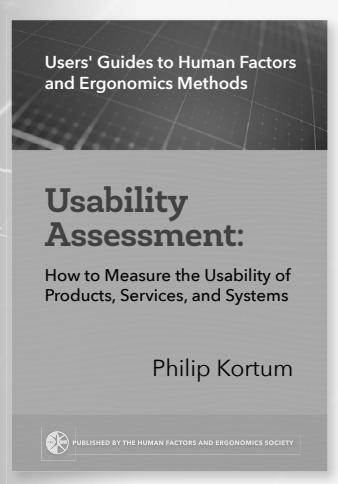


Table of Contents

1. What Is Usability Assessment?
2. Why Assess Usability?
3. Prepare to Perform the Usability Evaluation
4. Create Your Test Plan
5. Perform the Usability Test
6. Special Cases of Usability Assessment
7. Real-Life Example 1: Corporate Web Portal
8. Real-Life Example 2: High-Security Voting System
9. Some Parting Advice

This book will be valuable for undergraduate and graduate students; practitioners; usability professionals; human-computer interaction professionals; researchers in fields such as industrial design, industrial/organizational psychology, and computer science; and those working in a wide range of content domains, such as health care, transportation, product design, aerospace, and manufacturing.

ISBN 978-0-945289-49-4
120 pp., 7" x 10" paperback and e-book
http://www.hfes.org/publications/

PUBLISHED BY THE HUMAN FACTORS AND ERGONOMICS SOCIETY