Teamwork Evaluation with a Microworld Platform

Claudio Sapateiro
DSI–Polytechnique of Setúbal, Portugal
claudio.sapateiro@estsetubal.ips.pt

Pedro Antunes, Ijeoma Enwereuzo, David Johnstone
SIM–Victoria University of Wellington, New Zealand
{pedro.antunes, ijeoma.enwereuzo, david.johnstone}@vuw.ac.nz

José A. Pino
DCC–University of Chile, Chile
jpino@dcc.uchile.cl

***Abstract*—**We developed a platform supporting the evaluation of teamwork. The platform logs oral communications, collaborative tool usage and simulated physical action in the work environment. The paper brings forward observations derived from a comprehensive study of maintenance teams operating simulated network infrastructures. As positive results, we highlight that the platform supports a comprehensive analysis of teamwork at macro and micro levels, including individual and team activities, conducted in quasi naturalistic settings.

Keywords—Collaborative Systems, Microworlds, Evaluation.

I. INTRODUCTION

Our long-term research objective is studying how teams use technology to interact and collaborate. Teamwork is difficult to analyse. Firstly, it involves cognitive phenomena that are difficult to examine directly, such as situation awareness, sensemaking and decision-making [2]. Secondly, the phenomena are often entangled in multiple interaction patterns, which make it difficult to track events and detect causal relationships. Thirdly, typical data gathering instruments are either too intrusive to be effective or too detached from reality to be really insightful. Finally, we should consider that teamwork studies often may require the combined analysis of events within a spectrum of granularity, e.g., ranging from keystrokes up to decision-making. Developing research tools able to span such a wide range of events is a problem in itself.

Our research addresses these challenges by adopting a microworld approach to teamwork studies. Microworlds are real-time, task-oriented, synthetic environments used to study human behaviour in simulated, although intended quasi-naturalistic, scenarios [4]. One pertinent aspect of microworlds is that they combine a mix of characteristics of controlled and natural environments [7]. Research has been finding that microworlds are able to support naturalistic studies is spite of the constraints imposed by a semi controlled environment [10], as well as to promote learning and assessment associated with subtle skills, or allow experimentation on hardly accessible environments [12]. By providing some level of control, microworlds may help to uncover generalizable causal interpretations necessary to validate theory. As in the case of any technology and/or new practice introduction, impact assessment it will be more realistically evaluated closer to the very often multitasking environment in which it will take place instead of the solo foci on the specific intervention [13].

The developed Microworld platform systematises the teamwork evaluation process by addressing both interaction and collaboration, with consideration on the set of dependent and independent variables, accounting a controlled data gathering process. Next section reviews teamwork evaluation. The main requirements attended by the platform are presented in Sect. 3. Section 4 describes the platform. The platform use is discussed in Sect. 5. Finally, Sect. 6 gives concluding remarks.

II. OVERVIEW

As noted by [17], evaluating collaboration technology faces many practical, theoretical and methodological problems. Many shortcuts are often necessary to frame an evaluation process in a way that is at the same time efficient (from the researcher's point of view) and effective (considering the quality of the evaluation outcomes); thus multiple trade-offs have to be considered. We expand the [18] framework by identifying several dilemmas that should be considered when evaluating interaction and collaboration in teams.

Table 1 shows several dilemmas as semantic differentials highlighting a particular attitude towards evaluation. The first two dilemmas concern the classic distinctions between rigorous and relaxed experimental manipulations, and naturalistic and controlled settings. These two dilemmas have been studied by [18], who highlight the need to combine preliminary evaluations of early developments done in controlled settings with late evaluations of complete systems done in naturalistic settings. Dilemma No. 3 emphasises the researchers' goals when doing an evaluation, either seeking empirical or formal research.

Dilemma No. 4 emphasises that teamwork has individual and distributed dimensions, by their very nature requiring quite different theoretical scaffoldings. Finally, in dilemma No. 5 we acknowledge that understanding teamwork involves analysing both macro and micro activities.

Along with each dilemma we provide a list of advantages and drawbacks that can be found in related literature. Overall, what we observe is that every choice constrains the evaluation process in a different way, and considering those combined choices and their implications is certainly one of the major reasons making the teamwork evaluation a complex endeavour. This brings forward the potential value of microworlds. Microworlds, because of their semi open/closed nature, can combine laboratory experiments (rigorous, controlled) with field observations (naturalistic, minimal manipulation), thus resolving dilemmas 1 and 2. This semi open/closed nature comes from the support to unpredictable behaviour while strictly controlling and monitoring the participants' interactions [21]. For that reason microworlds have been referred to as "management flight simulators" [23], which highlights the degree of realism and engagement they can achieve [24]. Microworlds have already proved their utility in industrial process control [27], air traffic control [30], fire fighting [32], and other complex problem solving situations [34].

		Perspective and goals	Main advantages	Main problems
1	Rigorous manipulation	Gather information from controlled laboratory experiments; promote phenomena manifestations	Systematic validation of hypotheses and theory development	Relevance is highly constrained by the artificiality imposed by laboratorial settings [1]
	Minimal manipulation	Capture unconstrained information from the field	Openness and exploration	Experimental conditions may be hard to reproduce because of contextual dependency; prone to observe knotty phenomena
2	Controlled setting	Understand the relationships between dependent and independent variables	Eliminate confounding phenomena through isolation [3]	Cannot be applied to certain work contexts, e.g. high-risk situations [5, 6]
	Naturalistic setting	Understand how teams make decisions in real-world settings [8]	Eliminate the constraints imposed by laboratorial settings [9]	Dependence on practical problems related with task and context [11]
3	Behaviour	Validate theories and models explaining how humans behave [14, 15]	Complex processes can be analysed with methods such as process tracing and communication analysis [16]	Outcomes may not directly translate to technology development [14]
	Design	Gather pragmatic lessons from iterative design [14, 19]	Validation through utility assessment [19]	Outcomes rely more on common sense than fully articulated research hypothesis [20]
4	Individuals	Humans as information processing machines [22]	Studies of individual cognitive functions have been enriching the way we understand human behaviour [25, 26]	The whole is greater than the sum of its parts [28, 29]
	Teams	Expand our view from individuals to the relations between individuals and the environment where they operate[31]	Better/broadly correlate individual decisions to their actions [33]	Many complicating factors introduced by the aetiology of team's dynamics [35] and context [18]
5	Macro	Evaluate complex functions at a macro scale of performance [36, 37]	Research on micro phenomena lacks correspondence with the scale where teams perform complex tasks [39, 40].	Situated nature, dependent on concrete situations [3]; primary emphasis bears on experts [41]
	Micro	Understand how complex cognitive phenomena are entangled regarding to task execution	Allow precise control and measurement [3]; the more we reduce phenomena into elementary components the more general will be the principles [3]	Some cognitive phenomena are difficult to examine directly [2]; primary emphasis on routine tasks [41]

Table 1 - Evaluation dilemmas

Microworlds may also support the evaluation of technology designs, thus resolving dilemma 3. The key issues of dilemmas 4 and 5 concern the capacity to analyse teamwork at individual and team levels by gathering data with different granularity. Since microworlds usually mediate all team/user interactions, as with the environment and operational tools, they represent an ideal vehicle for overcoming the main problems raised by dilemmas 4 and 5. Next we derive some requirements from the previous dilemmas to inform Microworlds development. This is a pertinent contribution of this work, since one can mainly find in the literature microworlds applied within the scope of specific use cases and hardly a fine grain framework for orienting microworlds systematic development.

III. REQUIREMENTS

R1 - Control external events. This is related with experimental rigor and control addressed in dilemmas 1 and 2. The goal is balancing the teams' capacity to make decisions as if in a naturalistic setting with the capacity to capture behavioural data in a rigorous and controlled way. This involves controlling the injection of external events in the experimental scenarios, promoting context changes and unexpected reactions.

R2 - Mediate human-human, human-technology, and human-environment interactions. This requirement concerns dilemmas 2, 3 and 4. A key characteristic of teamwork is interaction; and the key goal of behavioural studies is examining interaction patterns. Three types of interaction can be considered: human-human (H-H), human-technology (H-T) and human-environment (H-E). H-H interaction involves information sharing, coordination and decision-making support, and other communicational based phenomena among humans. H-T interaction concerns the use of designed tools. Teams often use generic tools like social media software and shared editors; and they also use specialised tool designed to support the work domain. The interaction with these tools should be considered/captured by the platform.

H-E interaction considers the physical reality over which the team operate. Teams interact with the physical reality in various dimensions, work settings impose constraints that bound physical activities. An examples of such is the interaction with elements in the physical world such as mechanical levers. These interactions can be simulated by the platform in various ways. For example, adopting sophisticated immersion mechanisms to mimic the affordances of the real world, as seen in flight simulators. Other approaches with relaxed face validity may be considered depending on the

evaluation purposes and phenomena of interest [38]. In our research we only consider the latter case, further detailed below.

R2.1 - Human-human interactions. Interactions in the real world occur through different modalities, most often face-to-face, but radio, phone, chatting, e-mailing, and twitting are also common. The platform should reproduce the main characteristics of these modalities and in particular should preserve their one-to-one, one-to-many or many-to-many capabilities. Logging data according to these modalities is paramount because it affords information richness so necessary to analyse teamwork. Of course one interesting facet of studying teamwork is analysing the teams' communicational preferences according to context. But any comparisons must be done using a baseline. In our platform, the baseline is voice communication. We do not address other aspects found in face-to-face interactions like gestures and body language.

R2.2 - Human-technology interactions. As noted above, the platform should also support the evaluation of the envisioned operational technology design options. To accomplish this goal, the platform requires a model and interface of the technology being evaluated for integration.

R2.3 - Human-environment interactions. The platform considers two conceptual classes related to the physical environment: Location and Work Element. Locations are necessary to model teamwork done in multiple physical places, while Work Elements provide simulators for relevant physical interactions with the physical world, e.g. operating a physical machine.

R3 - Data logging must be contextualised at macro and micro levels. Considering that teamwork is open and dynamic, with multiple events injected over time and multiple interactions occurring in parallel, logs can be quite difficult to analyse. The problem is even more relevant when data is logged at both macro and micro levels. So an important requirement is keeping a coherent view of the relationships between the captured data and the environmental and task conditions triggered during the evaluation sessions.

R3.1 - Besides data logging, the platform should also support freeze-probes. Although logging users' interactions provides a large amount of data, in many studies that is not enough. Phenomena such as situation awareness, attention, stress, decision-making, and information overload can hardly be inferred from interaction data alone and thus require other ways of data gathering. Several approaches are then used, such as debriefings and talk-aloud protocols. The approach that seems in line with our perspective is using freeze-probes. The main idea is freezing momentarily the task and prompt users with some questions. Using freeze-probes in combination with logging allows dynamic generation of questions about the team's shared memory, awareness, workload perception and level of stress, etc.

IV. IMPLEMENTATION

A. Developed Platform

The implemented platform adopts a client-server architecture, the server provides data management, synchronization and logging, keeping a consistent environment and task state space representation; and clients mediating users participating in experimental sessions. What we call the client is indeed a set of four independent software components described below.

VoiceClient supports voice communication between team members. It implements the H-H interaction discussed in R2.1 (see III). In detail, VoiceClient can be configured to operate several communication channels in either unicast or multicast modes. Reproducing respectively the main features of phone/radio and conferencing calls. This component affords users to select the interlocutor by operating a simple control panel. VoiceClient is generic and can be reused across multiple evaluation processes and application domains.

TaskClient ultimately is the front-end that provides the considered operational possibilities mimicking real world affordances of the operational environment. It reflects and update the representational state space of the environment maintained in the server. For example, a team may work in various places, and understanding how physical distance affects teamwork may be valuable to the research. This is accomplished by the platform through the use of the Location abstract class (R2.3), to simulate the team members moving from a workplace to another (e.g. using a pull-down menu) and to convey the associated cost (e.g. taking time to physically move between places). It is also through Work Element (WE) abstract class (R2.3) that the simulator of any physical resource may be manipulated, through a set of planned attributes and operations, which establishes its state accordingly and propagate toward server State Space which holds the structure and dynamics of WE relationships.

However, identifying which task related features to implement depends on the case at hand and the phenomena of interest to researchers. This means that, unlike VoiceClient, the TaskClient may have to be developed for each application domain. Nevertheless, the abstract orientation followed on design and implementation decisions allow the partial reusability of the state space engine.

ToolClient aims to reproduce the functionality of a software tool(s) used by the team. For example, the functionality of a calendaring tool used by a team to coordinate their activities should be considered in the platform, as it necessarily impacts teamwork. The platform supports validating both fully-functional tools and early design ideas through simulation (R2.2). In both cases, ToolClient must provide a set of user-interface controls reproducing or interfacing tool functionality. ToolClient reusability is lower than TaskClient, since it has to be specifically developed.

FreezeProbeClient can periodically prompt participants about a set of task-related factors, e.g. situation awareness and shared memory (R3.1). FreezeProbeClient interacts with the server sate space to (optionally) suspend the on-going task and briefly question the participants. This component can be configured to collect various types of open and closed questions (e.g. yes/no, multiple choice) and thus it can be reused.

Table 2 highlights the controlled and naturalistic features of the platform components.

Tool	Naturalistic characteristics	Controlled characteristics	Reuse
Voice Client	Can reproduce typical functionality of mobile phones and walkie-talkies	User has to press different buttons to select a receiver and to start/stop voice communication	Yes
Task Client	Can reproduce specific actions in the physical space such as moving around and operating physical devices	Physical actions are substituted by check/set software functions, e.g. pressing a button instead of moving around	Partially
Tool Client	Can reproduce the user experience of a software tool	The interaction is detached from the physical device, e.g. the lab screen instead of a mobile device	No
Freeze Probe Client	Gather user data in context. Responses are not affected by hindsight and delays.	If not conceived with caution freeze-probes (which does not occur in the real world) may be prone to disruptions and bias	Yes

Table 2 - Characteristics of platform components

B. *Conducted Experiments*

We now delineate the whole evaluation process using the developed platform. We use as a demonstration a research study in the area of network maintenance (NM). NM teams ensure the operability of network components like computers, servers, and routers in large organisations. Critical Events in this domain are the loss of connectivity and device failures, which trigger multiple distributed decision-making activities to identify and solve the problems as soon as possible.

We built a TaskClient model to reproduce these characteristics, i.e. failure events and maintenance activities. Interactions with the network devices were modelled with a *check* operation, which tells if a device is working or not; and operational affordances such as *restart* or *replace* the device through Work Element instances which also couple the device status with the environment State Space representation on the Server. Of course the StateSpace accommodate impact chains of elements' state changes. In the currently described experiment is allowed for a failed router induce failures in a set of connected computers. Still related with the TaskClient, one characteristic of the physical domain that was also considered was the distance between network components. Rooted on an implementation of the Location abstract class users may virtually "move", e.g. from building B1, where computer C1 is located, to building B2, where router R2 can be found with an associated time cost.

The ToolClient for the NM scenario was developed with the aim to assess the design of a specific collaborative tool for urgent scenarios. The tool has been designed to support two functions: 1) assign tasks to individual team members (TMi), e.g. TM1 will check router R2 in building B2; and 2) report status of a network devices being checked/operated by a TMi, e.g. computer C1 in building B1 is working/malfunctioning after a restart. In our scenario, the team leader does task assignment. Task assignments and status reports are shared with all team members. Specific ControlPanels and ReportScreens were developed to deliver functionality through standard user-interface controls (e.g. buttons, menus, etc.)

The definition of freeze-probe questions, for the NM study, were rooted on the particular phenomena that drove that experiment: research the impact of the above mentioned tool on productivity and situation awareness. Productivity can be measured with a simple metric provided by the platform: time to successfully accomplish the exercise, i.e. identify faulty devices induced from injected faulty states and overcome them. Situation awareness related data was combined with log analysis and collected information from the users' freeze-probes.

Situation awareness can be defined as the capacity to know what the other team members are doing and overall scenario contextual state. In the NM study, we used the following (dynamic) questions to capture situation awareness at freeze probes: 1) What is the state of device X? 2) In what building is TMi? 3) Which device is currently causing the failure?

Shared situation awareness can then be operationalized as the sum of right answers to matching questions, given by every pair of respondents (i.e. if two users say that device X is working, and that status information is correct, then we add 1 to the situation awareness score). More considerations about measuring shared situation awareness can be found in [42].

The final thoughts over the platform use in the evaluation process concerned the experimental design. In our case, we were seeking to compare team performance when using/not using the ToolClient. Thus a repeated-experiments scenario was defined, so that half of the teams would complete a 15-minutes exercise while only exchanging information through the VoiceClient, followed by a 15-minutes exercise where they could use the VoiceClient and the ToolClient. The other half of the teams would complete the trials on a reverse order.

Multiple sessions with various participants, network configuration and different types of failures have been done using the platform. The obtained results indicated that the average time to complete the task was higher when using the ToolClient; that shared situation awareness was also higher when using the ToolClient; and also that the team preferred using the tool than communicating through voice [43]. Overall, the platform provided a large amount of data about the team behaviour, which could adequately answer several research questions about team behaviour. Though a detailed account of the research results from the NM study are outside the scope of this paper (details in [43]). Instead, we will discuss some lessons from using the platform as a research instrument.

V. DISCUSSION

The NM study discussed above highlighted some advantages and drawbacks of the evaluation platform. On the positive side, we may account for the following.

Capacity to gather a very large and diversified amount of experimental data. This included detailed logs of: 1) movements in the simulated physical space; 2) check/set operations done on network devices; 3) voice messages exchanged by the team; 4) actions done by users on the collaborative tool; 5) task assignments made by the group

leaders; 6) time necessary to complete the exercise; and 7) measure of each team's situation awareness. This diversity addresses dilemmas 4 and 5. The analysis contributed to conclude that the collaborative tool did not have impact on team performance (measured as time necessary to complete the task) but actually changed the teams' communication patterns by decreasing voice communications [44].

Capacity to engage teams in semi naturalistic scenarios. Of course the platform introduced some significant constraints to teamwork. For instance, it disallowed face-to-face communication and required teams to communicate through an uncharacteristic voice channel, which required them to press a few buttons to select a team member and start the communication, which is different from e.g. using a mobile phone. However, the platform did not impose significant constraints to other important features such as decision-making, both by team leaders (who have to assign tasks and continuously check what teams were doing) and other team members (who have to move between places and check/operate different devices).

Although we have been optimist about the capacity of the platform to reproduce some features of the naturalistic scenarios, we should also make some warnings. Our experiments highlighted that creating a semi naturalist scenario may require multiple iterations. For instance, some of our early experiments did not have any effort associated to moving between places. The result was that teams rapidly adopted a strategy where, instead of coordinating their activities (as they do in the real world), would rapidly and continuously move from one place to another to check the devices. These experiments had to be scraped from our study for lack of realism. Thus we have to carefully consider dilemma 2, observing that several iterations may be required to achieve an acceptable level of naturalness.

Capacity to reuse experimental instruments. In particular, we note the reuse of the communication and freeze-probe components. Our experiments underlined the advantages of suspending a collaborative task to inquire users about task and collaboration in context. Suspending teamwork can be difficult to achieve in truly naturalistic settings but is easy to orchestrate in the proposed platform, further considering the dynamics of the questions. Furthermore, users may be more promptly and effectively inquired than using other approaches such as debriefings and post-hoc analysis. This is especially important when data collection concerns fine-grained cognitive phenomena, such as group attention, task awareness, mental load, memory, impact of interruptions, etc. Overall, we observe that the proposed platform addresses dilemma 1 by combining repeatability with openness and exploration.

Combining behaviour analysis with design assessment. The platform supported controlled experiments, with treatments such as using a tool or not, which contributed to assess teamwork. In our study, the "no tool" treatment considered one-to-one voice communication, which was necessary to organise teams and report back the network device checks done by the team members. The "tool" treatment combined one-to-one voice communication with a coordination and information sharing tool. The data gathered from both treatments not only allowed detailed analysis of tool usage but actually revealed significant changes in communication patterns that could only be attributed to the tool. Since the platform supports repeating experiments with different tool design choices, we suggest the platform conveniently addresses dilemma 3.

On the negative side, we identify the following topics.

Model/software development. Software components had to be developed with data models needed to simulate the work environment and software tools of the teams. This effort is as much significant as the asked level of naturalness of the experiments.

Communication modalities. The approach requires mediating all communication through the platform, which of course does not support the richness of face-to-face interactions, although video can be added in the future.

Freeze-probing. As the research moves from micro to macro phenomena, the importance of logging users' activities diminishes while the importance of freeze-probes increases. However, developing adequate freeze-probes may not be an easy task. In our case, developing a freeze-probe for situation awareness was complex because various metrics could be considered mixing different appreciations for individual, shared and distributed awareness.

VI. CONCLUSIONS

All in all, the microworld strategy allows capturing insightful information about team collaboration combining quantitative and qualitative data, micro and macro phenomena, individual and team activities, some naturalistic and controlled activities, and also behavioural and design considerations. Our experiments with the developed platform point out towards the effective capacity to gather various types of measures, such as task efficiency, workload, communication load, and individual and shared awareness. Although these measures may not immediately lead to statistically significant differences, they contribute to explore research hypotheses.

We also recognise the possibilities that could be brought by using mixed groups in experimental research, where some team members may be real users while others may be artificially set up (possible to inform/constitute from trials such as this)[45]. This would further increase the value of the microworld platform outcomes (e.g. toward statistical significance).

We already mentioned the pros and cons of naturalistic and laboratorial approaches to the study of teamwork. We note the semi controlled/naturalistic approach advocated by our inquiry may represent a good trade-off between naturalistic studies and laboratory experiments. The developed platform can control variables related to teamwork apparatus while leaving other variables uncontrolled. The identification of which variables can/cannot be controlled and/or their mutual impacts may be very useful to research in this field. A deeper conceptual articulation of the proposed systematization for microworlds development and the Design Science research paradigm worth to be further explored.

REFERENCES

[1] R. Winter, "Design science research in Europe," *European Journal of Information Systems,* vol. 17, pp. 470-475, 2008.

[2] M. Endsley and D. Garland, *Situation Awareness Analysis and Measurement*. Boca Raton, FL: CRC Press, 2000.

[3] J. Flach, "Mind the Gap: A Skeptical View of Macrocognition," in *Naturalistic Decision Making and Macrocognition*, J. Schraagen, L. Militello, T. Ormerod, and R. Lipshitz, Eds.: Ashgate, 2008.
[4] R. Lew, R. Boring, and T. Ulrich, "A prototyping environment for research on human-machine interfaces in process control use of Microsoft WPF for microworld and distributed control system development," in *7th International Symposium on Resilient Control Systems*, 2014.
[5] B. Wallace and A. Ross, "Beyond Human Error - Taxonomies and Safety Science",, ed. T. Francis. 2006. ed New York: CRC Taylor and Francis Group, 2006.
[6] S. Dekker, *The Field Guide to Understanding Human Error*. Hampshire, England: Ashgate, 2006.
[7] W. Gray, "Simulated task environments: The Role of high-fidelity simulations, scaled worlds, synthetic environments, and laboratory tasks in basic and applied cognitive research," *Cognitive Science Quarterly,* pp. 205-207, 2002.
[8] G. Klein, "A recognition-primed decision (RPD) model of rapid decision making," in *Decision making in action: Models and methods*, G. Klein, J. Orasanu, R. Calderwood, and C. Zsambok, Eds. Norwood, CT: Ablex, 1993.
[9] E. Salas, S. Fiore, N. Warner, and M. Letsky, "Emerging multi-disciplinary theoretical perspectives in team cognition: An overview," *Theoretical Issues in Ergonomics Science,* vol. 11, pp. 245-249, 2010.
[10] G. Rolo and D. Diaz-Cabrera, "Decision-making processes evaluation using two methodologies: field and simulation techniques," *Theoretical Issues in Ergonomics Science,* vol. 6, pp. 35-48, 2005.
[11] R. Lipshitz, G. Klein, J. Orasanu, and E. Salas, "Taking Stock of Naturalistic Decision Making," *Journal of Behavioral Decision Making,* vol. 14, pp. 331-352, 2001.
[12] J. Gobert, Y. Kim, M. Pedro, M. Kennedy, and C. Betts, "Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld," *Thinking Skills and Creativity,* pp. 1–10, 2015.
[13] H. Hodgetts, S. Tremblay, B. Vallières, and F. Vachon, "Decision support and vulnerability to interruption in a dynamic multitasking environment," *International Journal of Human Computer Studies,* vol. 79, pp. 106–117, 2015.
[14] A. Hevner, S. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *Management Information Systems Quarterly,* vol. 28, pp. 75-105, 2004.
[15] A. Cleven, P. Gubler, and K. Hüner, "Design alternatives for the evaluation of design science research artifacts," in *Proceedings of the 4th international conference on Design science research in information systems and technology*, Philadelphia, PA, USA, 2009, pp. 1-8.
[16] J. Patrick and N. James, "Process tracing of complex cognitive work tasks," *Journal of Occupational and organizational Psychology,* vol. 77, pp. 259-280, 2004.
[17] P. Antunes, V. Herskovic, S. Ochoa, and J. Pino, "Structuring Dimensions for Collaborative Systems Evaluation," *ACM Computing Surveys,* vol. 44, 2012.
[18] D. Pinelle and C. Gutwin, "A review of groupware evaluations," in *Proceedings of 9th IEEE WETICE Infrastructure for Collaborative Enterprises*, 2000.
[19] K. Piirainen, R. Gonzalez, and G. Kolfschoten, "Quo Vadis, Design Science? – A Survey of Literature," in *Global Perspectives on Design Science Research*. Lecture Notes in Computer Science, vol. 6105: Springer, 2010, pp. 93-108.
[20] R. Briggs, "On theory-driven design and deployment of collaboration systems," *International Journal of Human-Computer Studies,* vol. 64, pp. 573-582, 2006.
[21] C. Gonzalez, P. Vanyukov, and M. Martin, "The use of microworlds to study dynamic decision making," *Computers in Human Behavior,* vol. 21, pp. 273-286, 2005.
[22] S. Card, T. Moran, and A. Newell, *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrance Elrbaum, 1983.
[23] J. Sterman, *Business dynamics: Systems thinking and modeling for a complex world*: McGraw-Hill, 2000.
[24] J. Keys, R. Fulmer, and S. Stumpf, "Microworlds and simuworlds: Practice fields for the learning organization," *Organizational Dynamics,* vol. 24, pp. 36-49, 1996.
[25] J. Reason, *Human Error*. Cambridge, UK: Cambridge University Press, 1990.
[26] P. Cacciabue, *Guide to Applying Human Factors Methods*. London: Springer, 2004.
[27] J. Sauer, D. Burkolter, A. Kluge, S. Ritzmann, and K. Schüler, "The effects of heuristic rule training on operator performance in a simulated process control environment," *Ergonomics,* vol. 51, pp. 953-967, 2008.
[28] N. Cooke, E. Salas, P. Kiekel, and B. Bell, "Advances in measuring team cognition," Arizona State University 2001.
[29] H. Cuevas, S. Fiore, B. Caldwell, and L. Strater, "Augmenting team cognition in human-automation teams performing in complex operational enviornments," *Aviation, Space, and Environmental Medicine,* vol. 78, 2007.
[30] K. O' Brien and D. O' Hare, "Situational awareness ability and cognitive skills training in a complex real-world task," *Ergonomics,* vol. 50, pp. 1064-1091, 2007.
[31] J. Gibson, *The senses considered as perceptual systems*. Boston: Houhton Mifflin, 1966.
[32] T. Chapman, T. Nettelbecka, M. Welsha, and V. Millsab, "Investigating the construct validity associated with microworld research: A comparison of performance under different management structures across expert and non-expert naturalistic decision-making groups," *Australian Journal of Psychology,* vol. 58, pp. 40-47, 2006.
[33] M. Turvey and R. Shawn, "Toward an ecological physics and physical psychology," in *The Science of the Mind: 2001 and Beyond*, R. Solso and S. Massaro, Eds. New York, NY, USA: Oxford University Press, 1995, pp. 144-169.
[34] M. Jobidon, R. Breton, R. Rousseau, and S. Tremblay, "Team response to workload transition: The role of team structure," in *Cognition: Beyond the brain: Embodied, situated and distributed congnition* Montréal, Canada, 2006, pp. 22-32.
[35] E. Salas, D. Sims, and C. Burke, "Is there a "Big Five" in Teamwork?," *Small Group Research,* vol. 36, pp. 555-599, 2005.
[36] F. Davis, "A technology acceptance model for empirically testing new end-user information systems : theory and results," Massachusetts Institute of Technology, 1986.
[37] B. Briggs, J. Nunamaker, and D. Tobey, "The technology transition model: A key to self-sustaining and growing communities of GSS users," in *Proceedings of the 34th Hawaii International Conference on System Sciences*, Hawaii, 2001.
[38] B. Brehmer, "Micro-worlds and the circular relation between people and their environment," *Theoretical Issues in Ergonomics Science,* vol. 6, pp. 73-93, 2005.
[39] P. Barnard, J. May, D. Duke, and D. Duce, "Systems, Interactions, and Macrotheory," *ACM Transactions on Computer-Human Interaction,* vol. 7, pp. 222-262, 2000.
[40] G. Klein, K. Ross, B. Moon, D. Klein, R. Hoffman, and E. Hollnagel, "Macrocognition," *IEEE Intelligent Systems,* vol. 18, pp. 81-85, 2003.
[41] S. M. Fiore, M. Rosen, K. Smith-Jentsch, E. Salas, M. Letsky, and N. Warner, "Toward an Understanding of Macrocognition in Teams: Predicting Processes in Complex Collaborative Contexts," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 52, pp. 203-224, 2010.
[42] P. Salmon, N. Stanton, G. Walker, and D. Jenkins, *Distributed Situation Awareness: Advances in theory, measurment and application to team work*: Ashgate, 2009.
[43] C. Sapateiro, "Evaluating Mobile Collaborative Applications Support of Teamwork in Critical Incidents Response Management," University of Lisbon, 2013.
[44] P. Antunes, C. Sapateiro, and J. Pino, "Supporting Experimental Collaborative Systems Evaluation," in *15th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Lausanne, Switzerland, 2011.
[45] V. Kurbalija, M. Ivanovic, C. von Bernstorff, J. Nachtwei, and H. Burkhard, "Matching observed with empirical reality-what you see is what you get?," *Fundamenta Informaticae,* vol. 129, pp. 133-147, 2014.