Human Error, Reliability, Resilience, and Performance, Vol. 173, 2025, 1–11
https://doi.org/10.54941/ahfe1006275

Virtual Human in the Loop (VHITL): Generating Synthetic Human Performance Data With HUNTER

Ronald Boring[1], Thomas Ulrich[2], Roger Lew[3], and Jooyoung Park[2]

[1]Scientitfic Visualization, Idaho National Laboratory, Idaho Falls, ID 83414, USA
[2]Human Factors & Reliability, Idaho National Laboratory, Idaho Falls, ID 83414, USA
[3]Virtual Technology and Design, University of Idaho, Moscow, ID 83844, USA

ABSTRACT

The Human Unimodel for Nuclear Technology to Enhance Reliability (HUNTER) software is used for dynamic human reliability analysis (HRA). HUNTER creates a digital human twin (or virtual operator) that interfaces with a digital twin (or nuclear power plant simulator). HUNTER is procedure driven, a unique characteristic of safety domains in which much decision making is rule based and captured in procedures. The outputs of HUNTER extend beyond the typical outputs of an HRA estimating method and approach the level of human performance data acquired from human-in-the-loop (HITL) studies using operators and a plant simulator. An advantage of HUNTER is that it creates a virtual human in the loop (VHITL). As such, HUNTER is a unique source of synthetic data on human performance. This paper highlights the use of HUNTER for use in automated evaluations for human factors. HUNTER augments HITL studies by providing a virtual tool to screen human interactions with novel technologies in the control room.

Keywords: Human in the loop, Virtual human in the loop, Dynamic human reliability analysis, Human performance data, Synthetic data

INTRODUCTION

As artificial intelligence (AI) technologies based in machine language or large language models evolve, the ability to use these tools for digital design and engineering tasks is similarly growing. An important element of design and engineering is considering the human user of those tools, which forms the basis of disciplines like human factors and user experience. These disciplines are still in their infancy in terms of using AI tools, especially in terms of optimizing user evaluations. This paper briefly explores opportunities how human factors can make use of simulation tools for evaluation of human performance. Following a brief exploration of why such tools are needed, we present a case study using a digital human twin from the nuclear power domain.

CONVENTIONAL HUMAN FACTORS

Definitions of human factors frequently mention it is the field of applying knowledge about humans to designing the products, systems, and environments they use (International Ergonomics Association, 2022). Much of the process of human factors as part of system design can be distilled into using different types of existing and gathered knowledge as evidence to inform the conceptualization and implementation of the system design. Human factors knowledge may be either *declarative*—meaning it encompasses those things we already know about humans—or *empirical*—meaning the human factors expert gathers new evidence about human performance. Declarative knowledge feeds into the initial phases of design to guide the conceptualization of the system, while empirical knowledge evaluates that the design actually works. This framework mirrors processes found in systems engineering, such as the popular Vee Model (International Council on Systems Engineering, 2023) as adapted in Figure 1. Alongside the Vee Model in Figure 1, the authors' trio of core human factors capabilities is depicted as they contribute to design, prototyping, and evaluation. Design corresponds to declarative knowledge, while evaluation is empirical knowledge. Prototyping bridges the two and enables empirical evaluations. The human factors process is often iterative, with feedback on early designs serving to refine designs that are repeatedly validated until eventual implementation.

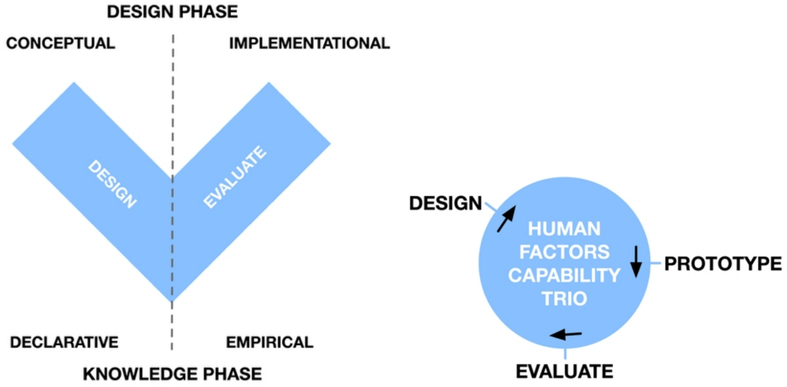

Figure 1: Design and knowledge phases overlaid on the systems engineering Vee Model (left) and the authors' human factors capability trio (right).

The Vee Model is named after its shape like the letter V, whereby the V represents verification and validation (V&V). V&V are standard evaluation processes employed in both systems engineering and human factors, representing different ways of pursuing empirical knowledge for the design (Boring, 2015). The *Guideline for Operational Nuclear Usability and Knowledge Elicitation* (GONUKE) framework offers insights into how V&V are used in practice in human factors (Boring et al., 2015 and 2021):

- *Verification:* A system that is being developed may be compared against criteria specified in a human factors checklist or standard.
- *Validation:* A system that is being developed may be tested through human-in-the-loop (HITL) approaches such as usability studies.

Verification, as defined here, harvests declarative knowledge by comparing the design to established criteria and best practices for system design. In contrast, validation gathers new knowledge of human performance while using the system. Verification uses the existing corpus of knowledge about good design, while validation is the sharp edge of novel design that builds the future corpus.

A crucial element of V&V is the specificity of knowledge required. As the design of the system matures from concept to implementation, the human factors burden of evidence shifts from general knowledge of principles of design to demonstration of actual use of the system. The eventual deployment of a system may require very explicit knowledge of the interaction of the human user with very specific use contexts of the system. Where declarative knowledge is insufficient, there is a need for new empirical knowledge to confirm the human-system interactions. The recent *Human Readiness Levels Standard* (Human Factors and Ergonomics Society, 2021) reinforces this principle—as the maturity of the technical system increases, it is necessary to ensure the human usability of that system tracks the maturity. It is necessary as the system matures to gather empirical knowledge of human performance to ensure the system is truly usable, safe, and human-ready for implementation.

A standard experimental psychology university curriculum will equip the psychologist with the skills to conduct empirical HITL studies. The experimental psychologist understands how to design a study to gather human data, a process that may involve learning how to manipulate conditions (i.e., the independent variables) and how to measure human performance (i.e., the dependent variables). Human performance measures may include common objective measures like time to complete the task, task completion accuracy, heart rate during specific tasks, and areas of visual fixation. Human performance measures may also include subjective measures like expert evaluations or self-reported measures like situation awareness or workload (Boring, 2015).

A specialization like human factors will add to this skillset specialized background knowledge on engineered systems and previous insights about humans using those systems. A human factors expert may specialize in areas like:

- Human errors in safety critical systems,
- User preferences for particular visual interfaces in control systems,
- Optimal processes and procedures for particular types of work tasks,
- Team communication in complex domains like healthcare, or
- Possible vigilance deficits while using automated systems.

Declarative knowledge (which is specialized for human factors) and empirical research skills (which are foundational to experimental psychology) combine to create a versatile researcher who can review and improve the human interface for almost any type of system. Where operating experience exists, this declarative knowledge can be harvested to shape the design of a new system. Where it does not exist, the researcher should conduct HITL studies to ensure that the system is usable by the human. The discipline of human factors contributes to the system design conceptually by establishing human-centered requirements and providing design guidance. Human factors also makes invaluable contributions to the system design implementationally by empirically evaluating human performance when using the system and providing recommendations to refine a design (see Figure 2). Declarative and empirical knowledge may be used along a formative-summative continuum, representing successive conceptual and implementation contributions to the design, or they may be used in an iterative manner within design phases.

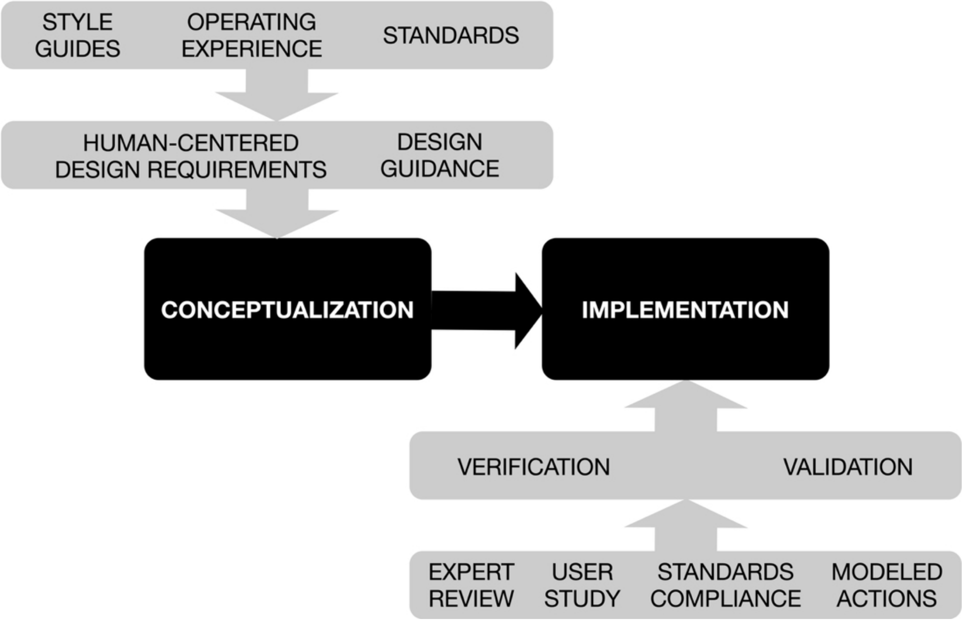

Figure 2: Conceptualization and implementation contributions to design by human factors.

AUTOMATED HUMAN FACTORS

As noted, verification is an expert-guided process of comparing a system against known human factors criteria, and validation is the process of conducting user studies. There is another possibility—using simulation instead of actual humans. Both verification and validation may be augmented by automated tools that eliminate the reliance on humans:

- *Virtual Verification:* Involves automating the task usually carried out by a human factors or subject matter expert and cross-walking to a standard. Rasmussen et al. (2017) allude to this in terms of a virtual analyst in human reliability analysis. Examples of this include automating usability evaluations for heuristic checklists, which have included rule-based (Xu et al., 2014) and machine-learning (Júnior et al., 2013) tools.
- *Virtual Validation:* Involves placing a simulated representation of the user into the scenario. For example, this could involve automatically

measuring ergonomic aspects like reach, fit, and comfort of a modeled environment using a digital mannequin (Institute for Energy Technology, 2024). Cognitive models have been used to test software usability (Wolf et al., 2018).

A comparison of human and virtual human V&V is found in Table 1.

Table 1: Human and virtual human verification and validation.

	Human	Virtual Human
Verification	Subject matter expert confirms system conforms to established human factors standards (e.g., heuristic evaluation)	Virtual analyst or expert system compares system to design standard (e.g., automated heuristic evaluation)
Validation	Representative sample of users interact with the system to produce human performance data (e.g., HITL or usability study)	Virtual human interacts with system to produce synthetic human performance data (e.g., digital mannequin across operational scenarios)

As noted by Abuaddous et al. (2022), much of the work on automating user experience and usability centers on tools to aid data collection for HITL studies, not actually to replace those HITL studies. To fill in this gap in research, here we explore the use case for virtual-human-in-the-loop (VHITL, pronounced "vittle") studies that do not use human participants. VHITL is still nascent, and virtual tools to develop user interfaces are much more mature than those for evaluation. Two recent developments in human reliability analysis (HRA) have paved the way for VHITL. These include the development of simulation-based HRA in the form of a virtual operator (or digital human twin), and the development of simplified simulators (or digital twins) that make initial proof of concepts possible.

EXAMPLE APPLICATION OF VIRTUAL HUMAN IN THE LOOP

Digital Human Twin: HUNTER

Traditional HRA methods are worksheet based, designed to analyze human error for specific event sequences. For example, a nuclear power plant risk assessment will include human actions (called human failure events) that can contribute to an abnormal event—either as human errors that exacerbate the event or as successful actions that lead to recovery. Dynamic HRA uses simulation to explore different outcomes of human interactions with the system. While conventional or static HRA methods make use humans to determine the human element of risk, dynamic HRA may omit the use of human analysts in determining the outcomes. Dynamic HRA employs resampling methods like Monte Carlo to model the range of human actions possible across plant scenarios.

The authors have developed the Human Unimodel for Nuclear Technology to Enhance Reliability (HUNTER; Boring et al., 2025b), a dynamic HRA method that creates a virtual operator model to walk through nuclear power plant scenarios and mimic human actions. The basic HUNTER framework consists of three modules:

- *Individual Module:* Models the factors that improve or degrade human performance,
- *Task Module:* Models the actions the human performs as contained in the plant procedures, and
- *Environment Module:* Models the plant in which the human operates, which is handled by the plant simulator.

These modules are essential to any human simulation. The nuclear power context is ideal for dynamic HRA, because plant operations within the main control room are heavily specified in the form of written operating procedures, and every plant is required to have a high-fidelity simulator for training. Thus, HUNTER is a type of procedure-based automation system. It replicates human operators by following procedures within the simulator. However, unlike an actual plant automation control system, HUNTER is designed to encompass a range of behaviors, from optimal following of the procedures to contexts where human error might manifest. For example, in a situation where an operator is highly stressed, the operator may perform tasks more quickly or slowly than normal and be more prone to skipping steps in the procedure. This type of human performance variability is captured using HUNTER's Individual Module. Overall, the Individual and Task Modules comprise the digital human twin or virtual operator, while the Environment Module in the form of the simulator is the digital twin (Boring et al., 2023a).

Digital Twin: Rancor

The Rancor Microworld Simulator (Ulrich et al., 2017) was originally designed as a simplified simulator to aid in the collection of HRA data. With human error probabilities for most main control room actions predicted to be 1/100 or less, small scale simulator studies involving a single crew do not provide adequate opportunity to observe human errors organically (Medema, Boring, and Mohon, 2021). Rancor provides a simplified plant model with a digital human-machine interface. In contrast to most training simulators for nuclear power plants, Rancor can be learned quickly, allowing studies to be conducted with students or other novice populations. Benchmark studies have demonstrated that the performance of student operators using Rancor is analogous to professional reactor operators using higher fidelity simulators (Park et al., 2023). Given the generalizability of these results, it makes Rancor an ideal platform for gathering human performance data to inform HRA, because large sample sizes suitable for HRA may be easily obtained.

Rancor includes the Rancor Integrated Procedure System (RIPS; Boring et al., 2025a), which is a computer-based procedure system that allows full control of the simulator while also automating the heretofore arduous task of logging human operator performance with other plant parameters. Capturing operator performance data opens up considerable opportunity to use HITL studies with Rancor for HRA or machine learning purposes (Boring et al., 2022).

A unique feature of RIPS is its ability to interface with HUNTER as the Task Module. HUNTER uses RIPS to gather information about the plant and take control actions virtually. A recent feature addition to RIPS is the ability to interface with plant simulators beyond Rancor, thereby increasing the potential of HUNTER to run realistic simulations of operators from actual plants. The HITL data gathered using RIPS can be used to calibrate HUNTER model performance to match actual operator performance.

VIRTUAL ASSEMENT OF HUMAN ACTIONS

The combination of HUNTER and Rancor enable first-of-a-kind VHITL studies. There are two main use cases for VHITL:

- *Plant Modernization:* Existing nuclear power plants are undergoing significant upgrades from legacy analog instrumentation and controls to digital control systems. In many cases, these digital upgrades require changes in the operation of the plant, including the development of new procedures. Previously demonstrated functionality in the form of HUNTER-Procedure Performance Predictor (P3) allows procedure developers to stress-test new procedures and identify problems with procedures as written or the procedures as used by the operators (Boring et al., 2023b).
- *New Builds:* Existing nuclear power plants have mature risk assessment models. The need for new HRA methods may be questionable, when the current methods already adequately model human actions. Adopting new methods would require considerable rework of risk models. However, new plants being designed and built will feature many new technologies and concepts of operations, presenting a significantly different risk profile than existing reactors (Boring, 2023). Dynamic HRA methods like HUNTER are well suited for these ground-up risk analyses. Specifically, since there is not always a legacy of operating experience to draw on for these new plants, the ability to simulate operator performance aids in anticipating problem areas for human activities.

To this latter point, Virtual Assessment of Human Actions (VAHA) is a framework by which a virtual operator may be used to support both HRA and human factors empirical data needs. VAHA makes use of the digital human twin concept to run through scenarios with novel human-technology interactions. Of particular concern to both HRA and human factors is that the range of these interactions may not be well understood. For example, a novel control system that features automation may have an expansive range of emergent interactions that are unanticipated and unexampled. Testing and validating these interactions involves sampling scenarios for HITL studies. It is simply not possible to sample all possible scenarios, meaning it may never actually possible to validate a novel system completely through HITL. It is conceivable that there will remain gaps in testing that could prove risk-important. Existing methods of HITL testing push the limits of engineering, which may not be able to anticipate all use contexts. If misrepresentative or insufficient scenarios are tested through HITL, the data are simply inadequate to validate the design.

VAHA helps reduce the possibility of validation gaps. Where it is not possible to expertly anticipate every use context of a new system, simulation techniques allow screening a wider range of interactions than would be possible using HITL alone. VHITL using a tool like HUNTER can easily run through thousands of permutations of the human operator interfacing with the novel system, flagging those contexts that result in human errors or otherwise degrade system functioning. VHITL does not replace HITL; it complements it (see Figure 3). VHITL can run through the gamut of scenarios and interactions, thereby screening problematic human activities. Where VHITL identifies issues, these can be reviewed and used to select scenarios that may be of interest with actual HITL studies. HITL remains the gold standard for evaluation, but VHITL supplements it by down-selecting those human actions of interest for more extensive human user testing.

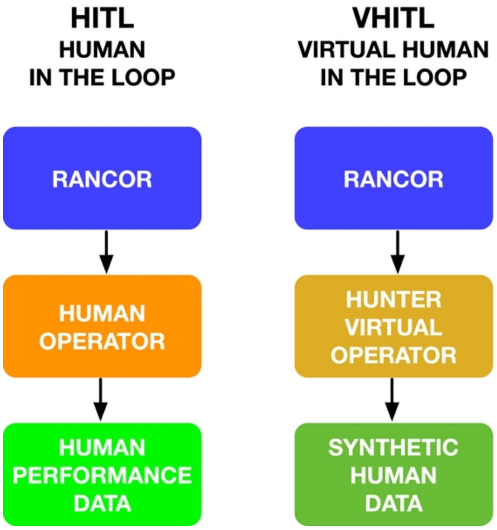

Figure 3: Human and virtual human in the loop process prescribed in VAHA.

Of course, the utility of VAHA is predicated on having an accurate simulation of human performance. Within the realm of nuclear power operations, HUNTER has been found to produce a faithful reflection of actual human performance for a variety of normal and off-normal scenarios. The ability to generalize to new scenarios and domains is the focus of ongoing research.

VAHA benefits both HRA and human factors. The traditional application of HRA is to provide human error probabilities for important human actions, which feed into the overall plant risk assessment. Dynamic HRA methods like HUNTER can produce many more outputs beyond error probabilities. For example, HUNTER may be used to calculate task durations—an important consideration for time-critical tasks. Success or failure may not hinge of the absence or presence of overt human errors; it may depend on the ability to complete a task within a prescribed time. Additionally, HUNTER may be used to estimate common human factors measures like situation awareness and workload. Degraded states of performance may not rise to the level of human error but are definitely of interest to the human factors assessment of the system.

CONCLUSION

HUNTER's synthetic data for human performance augment human performance data derived from HITL studies. They represent an important early implementation of automated evaluation techniques. Following the VAHA approach, VHITL can be used as a screening method not just to estimate human error probabilities but more broadly to anticipate the types of human interactions that will occur with novel technologies. This approach—when combined with emerging AI-powered tools for digital design and engineering—promises to accelerate both the conceptualization and implementation phases of system development. In an era of increasing electricity demand and the push for deployment of more nuclear power plants to meet that need, automated evaluation is an essential tool toward licensing new plant designs.

Currently, HUNTER is not based in the AI areas of machine learning nor large language models. Rather, it represents a production system that follows a scripted path, with flexible outcomes. The approach is a psychology-based model akin to a physics-based model, not a data-driven model common in machine-learning research. This approach works well in a highly proceduralized area like nuclear power control room operations, but its scalability to other domains may be limited. The authors are exploring approaches beyond existing HUNTER that would allow the use of VHITL as a method and VAHA as a framework to be used in more areas. Such techniques include using Rancor human and HUNTER synthetic performance data as training data for machine learning applications and exercising procedures as a corpus for a large language model. Certainly additional frameworks for digital human twins beyond HUNTER should be developed to allow virtual evaluations. The future of design and evaluation will require many automated tools, and virtual humans will be essential for the completeness of HRA and human factors evaluations of complex systems.

ACKNOWLEDGMENT

This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory, an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately-owned rights. Idaho National Laboratory is a multi-program laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517.

REFERENCES

Abuaddous, H. Y., Saleh, A. M., Enaizan, O., Ghabban, F., & Al-Badareen, A. B. (2022). Automated user experience (UX) testing for mobile application: Strengths and limitation. International Journal of Interactive Mobile Technologies, 16(4), 30–45.

Boring, R. L. (2015). Envy in V&V: An opinion piece on new directions for verification and validation in nuclear power plants. Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society, 59, 1746–1750.

Boring, R. (2023). Human factors for advanced reactors. AHFE Open Access, 94, 156–165.

Boring, R. L., Lew, R., & Ulrich, T. A. (2025a, in press). Rancor Integrated Procedure System (RIPS): A computer-based procedure platform for advanced reactor research. Proceedings of the 14th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies.

Boring, R., Mortenson, T., Ulrich, T., & Lew, R. (2022). Humans with/as big data in nuclear energy. AHFE Open Access, 54, 56–64.

Boring, R. L., Ulrich, T. A., Joe, J. C., & Lew, R. T. (2015). Guideline for operational nuclear usability and knowledge elicitation (GONUKE). Procedia Manufacturing, 3, 1327–1334.

Boring, R. L., Ulrich, T. A., & Lew, R. (2021). Putting GONUKE into practice: Considerations for human factors evaluations. Proceedings of the International Annual Meeting of the Human Factors and Ergonomics Society, 65, 618–622.

Boring, R., Ulrich, T., Lew, R., & Park, J. (2023a). Synchronous vs. asynchronous coupling in the HUNTER dynamic human reliability analysis framework. AHFE Open Access, 82, 43–52.

Boring, Ronald, Ulrich, Thomas, Lew, Roger, Park, Jooyoung, (2023b). HUNTER procedure performance predictor: Supporting new procedure development with a dynamic human reliability analysis method. AHFE Open Access, 117, 29–38.

Boring, R. L., Ulrich, T. A., Lew, R., & Park, J. (2025b, in press). Rancor-HUNTER: Using a simulator engine for realistic human performance modeling of nuclear power operations. Proceedings of the Probabilistic Safety Assessment Conference.

Human Factors and Ergonomics Society. (2021). Human Readiness Level Scale in the System Development Process, ANSI/HFES 400–2021. Washington, DC: Human Factors and Ergonomics Society.

Institute for Energy Technology. (2024). Halden Virtual Reality Centre Create Overview, Release 4.2.

International Council on Systems Engineering. (2023). INCOSE Systems Engineering Handbook. A Comprehensive Reference on the Discipline of Systems Engineering. 5th Edition. Wiley.

International Ergonomics Association. (2022). Giving Your Business the Human Factors Edge. Geneva: International Ergonomics Association.

Júnior, D. G. S., Hernández-Ramírez, R., Estima, J. (2024). Investigating usability indicators for the adoption of AI models in heuristic evaluation. Springer Series in Design and Innovation, 35, 136–151.

Medema, H. D., Boring, R. L., & Mohon, J. D. (2021). Extracting human reliability findings from human factors studies in the Human Systems Simulation Laboratory. Proceedings of the 2021 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA 2021), 98–107.

Park, J., Yang, T., Boring, R. L., Ulrich, T. A., & Kim, J. (2023). Analysis of human performance differences between students and operators when using the Rancor Microworld simulator. Annals of Nuclear Energy, 180, Article 109502.

Rasmussen, M., Boring, R. L., Ulrich, T., & Ewing, S. (2017). The virtual human reliability analyst. Advances in Intelligent Systems and Computing, 589, 250–260.

Ulrich, T. A., Lew, R., Werner, S., & Boring, R. L. (2017). Rancor: A gamified microworld nuclear power plant simulation for engineering psychology research and process control applications. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61, 398–402.

Wolf, K. I., Thalappully, R., Wagner, Y., Wallhoff, F., & Appell, J.-E. (2018). Novie2Expert—A cognitive model within a usability evaluation framework. 7th Interdisciplinary Workshop on Cognitive Systems.

Xu, J., Ding, X., Huang, K., & Chen, G. (2014). A pilot study of an inspection framework for automated usability guideline reviews of mobile health applications. Proceedings of the Wireless Health 2014 on National Institutes of Health (WH '14).