Using Rapid-Play Serious Games and Crowdsourcing to Inform Technology Evaluation and Research Prioritization

Robert M. Seater

Rapid-play serious games can allow players to gain intuition about the use of a proposed capability, enable researchers to examine that capability's influence on tactics and procedures, and collect quantitative data that supplement qualitative user feedback to inform decisions about which new technologies should be pursued with future development.

The analysis of user-facing future technology is a difficult task but one that plays an important role in the process of research, development, and technology evaluation (RDTE). The RDTE process includes many facets, ranging from brainstorming potential threats and opportunities all the way to prototyping and conducting field evaluations. An efficient RDTE process is important to avoid missing opportunities (culling good ideas) or investing too much effort into dead ends (failing to cull bad ideas). Unfortunately, many technology programs fail before they even get started because they are seeking to provide a capability that users do not need or will not accept. However, recognizing which technologies will be useful before they have been developed, prototyped, and field tested can appear to be a chicken-and-egg problem—how can we triage a set of capabilities before they exist?

To understand how to address this problem, it is first important to articulate what makes the task difficult. Consider, for example, a proposal for a novel detection technology that is light enough to be used as a wearable sensor for infantry squads. If it is our job to decide if that technology is worth maturing for that application, we face several immediate challenges:

- First of all, because the technology does not exist yet, we don't know what technical tradeoffs it will be able to offer, what technical specifications we would want it to meet, or where additional research is most needed to close the gap. Is it more important that the sensor have a low false-positive rate or a high range? A high-fidelity image or a fast update rate? We don't even know where a research program should focus its efforts or if the end result will be acceptable to users.

- To answer such questions, one typically turns to current domain experts and users. Involving experts and users can provide valuable feedback on the utility of the new capability and its likelihood of being accepted. So, we might ask current squad soldiers what they would find most helpful in a wearable sensor. Unfortunately, most expert decision makers are intuitive thinkers used to dealing with concrete situations, not abstract thinkers who have a theoretical formalism that can generalize to future scenarios [1]. Expert users may not understand why they are experts and thus not understand what new capabilities will help them in a novel (future) environment [2].

- To make the problem more concrete for the domain experts, we might run a tabletop exercise or seminar-style wargame [3] so that they can get some intuition for what it is like to use the proposed capability and how it might change their operating environment. However, after such an exercise (or even a few), the domain users are still novices at using the new technology, and they haven't had much chance to experiment with how to use the technology in different ways or to explore how it might change doctrine and best practice. The squad members have only had a couple of chances to experience how a wearable sensor might change their behavior and how to incorporate it into current doctrine. In an adversarial setting, the red force will also not have had time to develop exploits and counter-tactics. Furthermore, we still rely on participants' qualitative descriptions of what they liked or didn't like about using the sensor—a method hindered by users with dominant personalities or experts who are not good at theorizing.

- To address the issues that come from a small number of qualitative data points, we might run a large number of exercises and instrument users to collect data on their performance and behaviors. However, that is an expensive proposition if one uses traditional exercises and tabletop scenarios that take hours or days to run, that pull experts away from other tasks, and that require participants to travel to a common location. Such an approach is costly, burdensome, and slow. The early phases of RDTE can seldom afford any of those drawbacks, and developers usually face pressure to provide a quick, cheap, and low-burden estimate of where to focus subsequent efforts so that the next phase of the program can get under way with most of its budget intact. If we spend all our time understanding what wearable sensor to build, the program may be canceled or the problem may simply become obsolete as the world changes.

So what we are looking for is a method of providing users with a concrete environment in which they can explore a future capability many times to build intuition, collect both quantitative and qualitative data on their performance and preferences, and do so without consuming a lot of program time, participant time, or budget.

HIVELET: Crowdsourcing Human Creativity

For the last few years, MIT Lincoln Laboratory has been using serious games to aid in technology assessment programs. One of the most recent efforts is the Human-Interactive Virtual Exploration for Low-Burden Evaluation of Technologies (HIVELET). The HIVELET approach focuses on early RDTE, especially when suites of emerging technology are being considered for user-facing roles. This approach combines economic game theory [4] with rapid-play digital simulations to collect quantitative data, improve qualitative feedback, and crowdsource the ingenuity of human experts.

Under the HIVELET approach, players alternate between two modes—capability selection and mission simulation, as illustrated in Figure 1.

- *Capability selection* allows players freedom to select different combinations of conceived capabilities, allowing them to formulate and explore different strategies that may

deviate from current doctrine. However, the selection mode prevents a player from simply choosing all available capabilities; they must manage a limited budget (representing cost or weight), forcing them to think critically about what capabilities they really need and to carefully prioritize the available capabilities. Players are not only judging if a capability is useful but also if it is useful *enough*, given its drawbacks and alternatives.

- *Mission simulation* gives the player a chance to try out the set of capabilities they selected to get feedback about their effectiveness and to build intuition about what did or did not work well. The mission simulation is focused on being short (e.g., minutes not hours) so that the player can make multiple attempts within a single sitting to explore different strategies and build more intuition through iteration. To achieve these objectives, the mission simulator captures a key aspect of a critical decision point in the real-world and abstracts away details not relevant to the evaluation at hand. Design principles and scoring incentives are used to create an environment that accurately recreates the pressures of the real world while simplifying the real-world simulation enough to shorten the duration of gameplay.

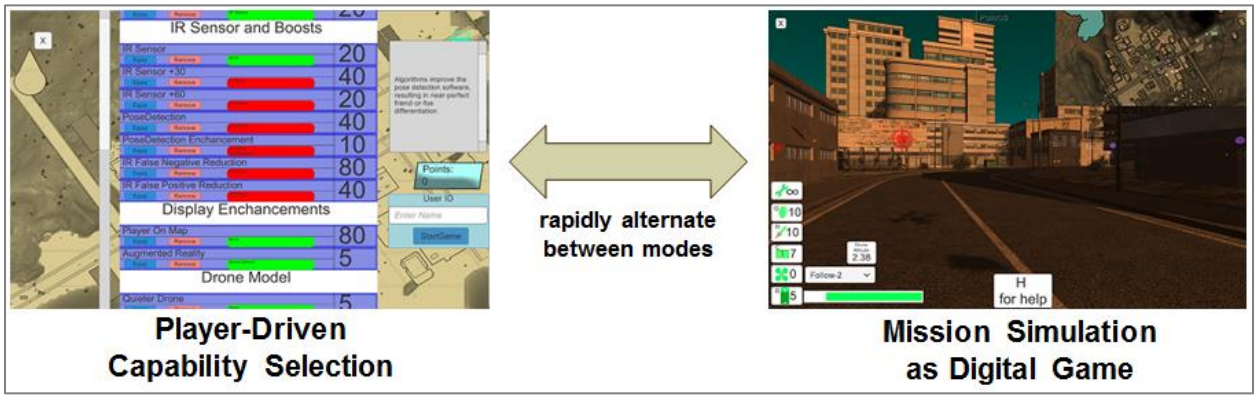

FIGURE 1. Under the HIVELET approach, players alternate between two modes—capability selection (left) and mission simulation (right). By rapidly alternating between the two modes, players can build intuition for future capabilities, explore novel uses of the capabilities, and allow the researchers to collect data on players' preferences, behaviors, and performance.

After completing the mission simulation using the selected capabilities, players return to the selection mode. They can stick with their prior choices, refine their strategy, or try an entirely different approach. They then repeat the simulation, continuing to alternate back and forth between the two modes. The alternation forces the player to combine abstract thinking about the value of various capability combinations with concrete feedback and intuition about the use those capabilities on a mission. Data collected during the game reveal the player's preferences, behaviors, and performance and can be used in researchers' quantitative analyses that complement the qualitative feedback provided by participants. With appropriate design of the framework, a participant can complete several cycles of selection and simulation in an hour.

Both portions of the game can be hosted online and played remotely by participants, thereby greatly reducing the burden and cost per each data point. A wide range of players remotely playing a series of short simulations can quickly compile a lot of data that can shed light on the tradeoffs and priorities for the capabilities being modeled. Researchers can also vary the mission parameters to see how players change their preferences and strategies, thereby providing insight into the application or the concept of operation (CONOPS) for which a given future capability is likely to be best suited. For example, the infantry mission simulator shown in Figure 2 can be run using a range of different terrain types and mission objectives to determine the flexibility or specialization of certain capabilities.

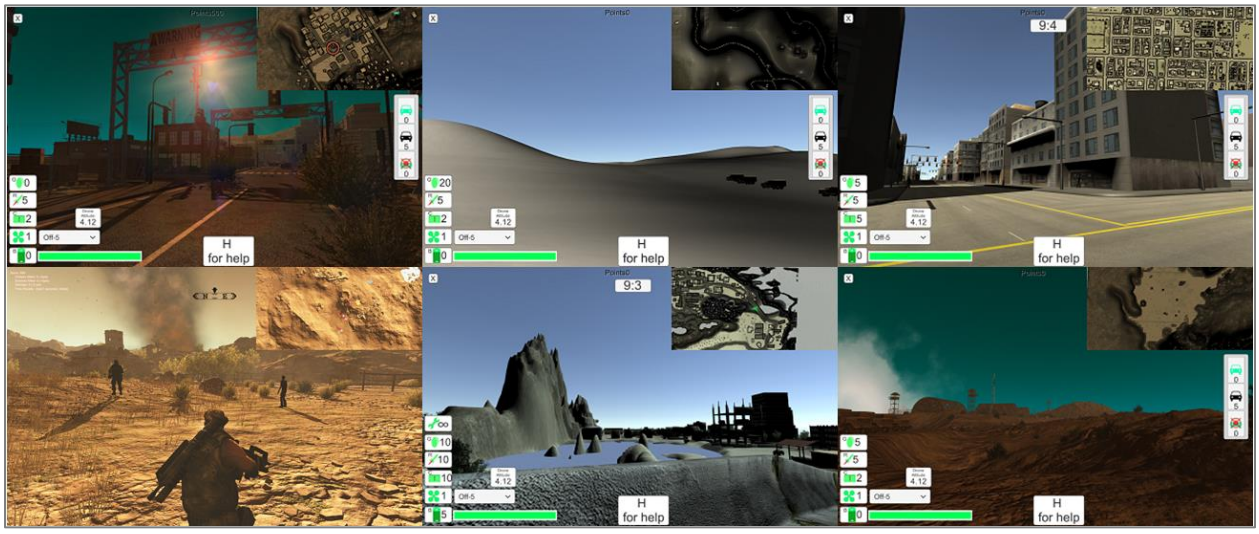

FIGURE 2. A player executes a tactical infantry mission in a digital simulation, using in-game models of concept technologies. Domain experts who rely on experience and intuition often find it easier to provide feedback on concepts when they can try them out in a simple simulation rather than when they are asked to engage in a purely theoretical discussion. Researchers can examine how player behavior and preferences change in different environments and for different missions. The environments shown here are a ruined city, an arctic tundra, a large city, a rocky desert, an island, and a night mission.

This approach is a form of crowdsourcing—using humans in large numbers to perform tasks that are difficult to automate. In this case, the task being automated is the creative thinking and ingenuity about how to mix and match future capabilities of various quality levels into a coherent and effective strategy that manages the risks presented by a real-world mission situation. Humans are not good at fine-tuned optimization, but they are excellent at creatively finding good combinations from within a very large decision space. This approach is thus well suited to the early stages of RDTE, in which we need to rapidly triage an enormous design space to focus more systematic traditional evaluation methods on the most promising options. HIVELET isn't the end of the RDTE story, but it can be a critical step in making other techniques more focused, more efficient, and ultimately more likely to succeed than they would be if used alone.

Application to Infantry Technologies

The HIVELET technique is a new approach that is still undergoing validation and development, but the early results are very promising. It has been used to evaluate how a small unmanned aerial vehicle (UAV) integrated into tactical infantry missions might fundamentally change how such squads operate. The game modeled 29 capabilities (e.g., sensors and control mechanisms) and capability upgrades (e.g., enhancements to the sensor quality or to the player's weaponry). In the mission simulator, players navigated a three-dimensional (3D) real-time environment and attempted to recover data from a predator drone that went down in a hostile urban environment (Figure 3). The player has to balance finding the objective quickly with safely navigating the terrain to avoid or neutralize threats.

FIGURE 3. In this mission simulator, players must navigate a hostile urban environment to find a crashed predator drone, recover its data, and extract those data safely. They must choose between future capabilities that improve the efficiency of the mission and the safety of their squad, and are encouraged to experiment with nonstandard tactics enabled by those capabilities.

HIVELET supports a range of different selection mechanisms (drawn from economic game theory) that impose different limitations on what capabilities players can bring on each mission. Different selection mechanisms can be useful for collecting different types of data. In this application, the players used a random market, i.e., before each mission, the players are presented with a list of all available capabilities, each of which has been assigned a random price, as shown in Figure 4. They may select any number of those capabilities, but the prices are deducted from their upcoming mission score. In this manner, players are pressured to make do with as few upgrades as possible, driving them to think critically about the relative values of different capabilities.

The capabilities available included radio-frequency (RF) sensors that help locate the objective, infrared (IR) sensors that help identify potential hostiles, image processors that help differentiate civilians from hostiles, various control mechanisms for the personal drone, user interface displays available to display sensor data, and advanced munitions to give the player improved firepower. Players could combine these capabilities to support a range of strategies, both conventional and unconventional. For example, players might buy a "follow-me" control mechanism, an IR sensor, and an augmented-reality helmet display, then perform the mission on foot with a visual indicator of nearby potential threats (such a strategy is depicted in use in Figure 5). Alternatively, they could buy an onboard camera for their UAV, robotic underarms, and an onboard RF sensor, then attempt to find the objective and complete the mission entirely with the drone, without putting their own characters at risk.

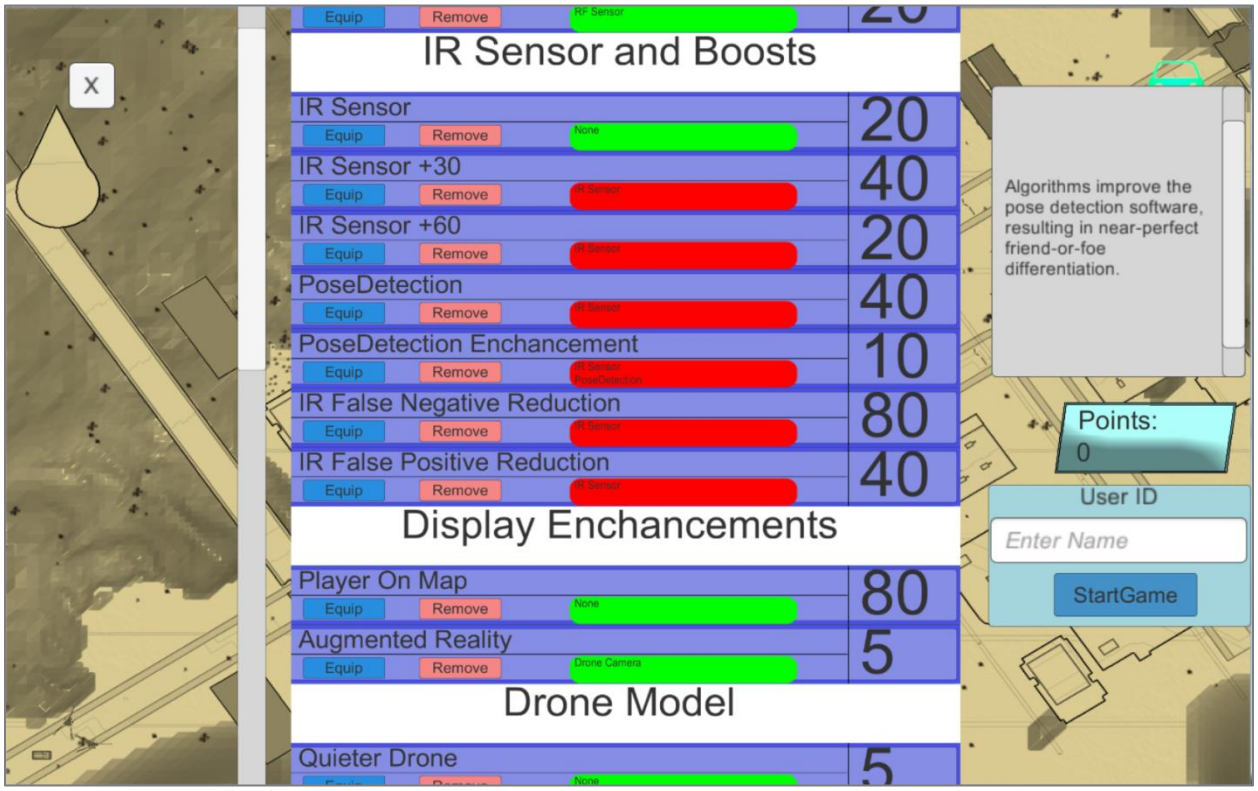

FIGURE 4. The technology selection screen used by players to choose what capabilities they will combine for the next mission. Each capability has a cost, forcing the player to make critical decisions about their priorities, while allowing them to explore different combinations and discover novel strategies.

FIGURE 5. After selecting capabilities, players try them out in a real-time simulation of an infantry mission, allowing the players to experiment with new strategies and gain intuition for technology utility, and enabling researchers to collect data on players' choices and behaviors.

Bringing Quantitative Analysis to Early Concept Analysis

Much of the work thus far on HIVELET has been on validating its merit rather than on applying its technique to particular domains. Data collected from initial experiments indicate that the technique is capable of quickly providing useful quantitative data about the value of future technologies. In this section, we review some of the quantitative analyses that are enabled by this style of rapid-play serious game.

We can understand the value of using rapid-play serious games by looking at data collected from users who are interacting with the system, including participants with a mix of research and military backgrounds. From the data collected about player choices, performance, and behaviors, we can see that the technique is able to bring data analytics to bear on answering questions about future technology. Figure 6 shows that a few hours of gameplay is sufficient for players to start providing coherent data to be analyzed: 1 hour of training plus 1 hour of solo play was enough for players to stabilize their scores and start producing consistent levels of performance. Players' scores were calculated from a combination of completing the mission, avoiding enemy fire, and minimizing the number of technologies purchased. Players self-reported that 1 to 2 hours of exposure was sufficient to learn the game, formulate a strategy,

execute the strategy, and develop opinions about the value of the technologies, at least within the context of the mission simulated in the game.

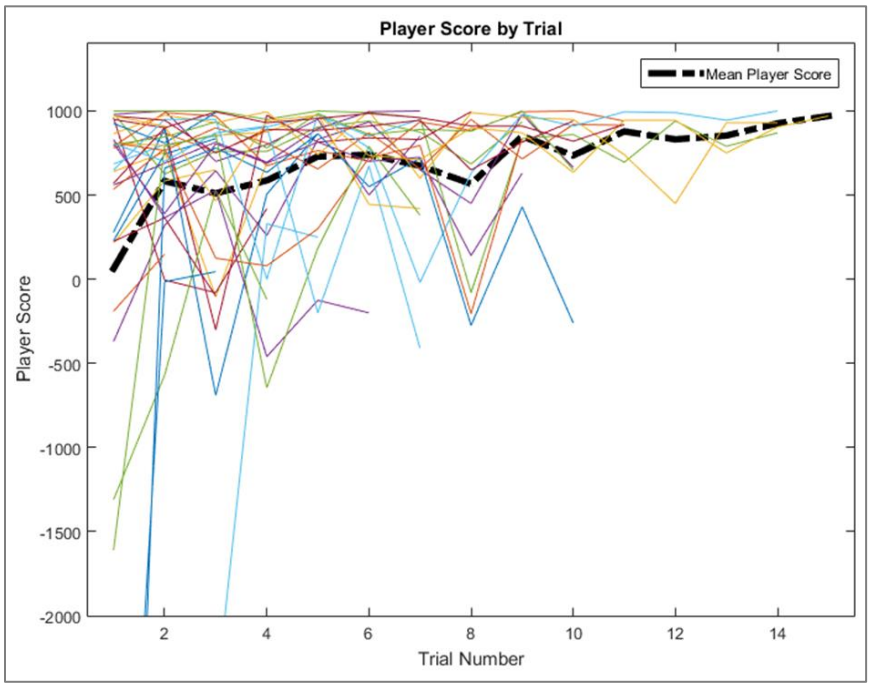

FIGURE 6. Players' scores improved and converged over the course of a 1-hour play session. Participants had never seen the game prior to this trial other than for a 1-hour training session. Within 1 to 2 hours of gameplay, players were able to learn the game, formulate a strategy, execute the strategy, and develop opinions about the utility of modeled future technologies.

Once we believe that players have had sufficient time to develop opinions, we can examine what values they expressed. Figure 7 shows the frequency with which each of the 29 modeled technologies was selected over all participants, and we can see strong trends in player preferences within this mission context – finding a crashed airborne asset in a hostile urban environment. Drone-mounted cameras and long range drone-mounted radio-frequency sensors were highly valued, as they allowed players to quickly and safely scout for the lost asset. Interestingly, short range drone-mounted radio-frequency sensors were considered to be almost useless, which helps us to establish the minimum acceptable requirements for such a device.

Drone-mounted IR sensors of any range were selected very rarely by players. This result initially surprised the research team as the IR sensors allowed players to know where hostile forces were in the city. This valuation makes more sense when paired with the qualitative feedback from players, who described the best strategy as running the entire mission with the personal drone and avoiding ever entering the city on foot. Thus, knowing the location of hostile forces was not important to this mission given the available technologies, and players discovered a strategy not anticipated by the research team. One of the strengths of rapid-play

games is their ability to allow players to experiment with new strategies and anticipate how future technology will change tactics and doctrine.

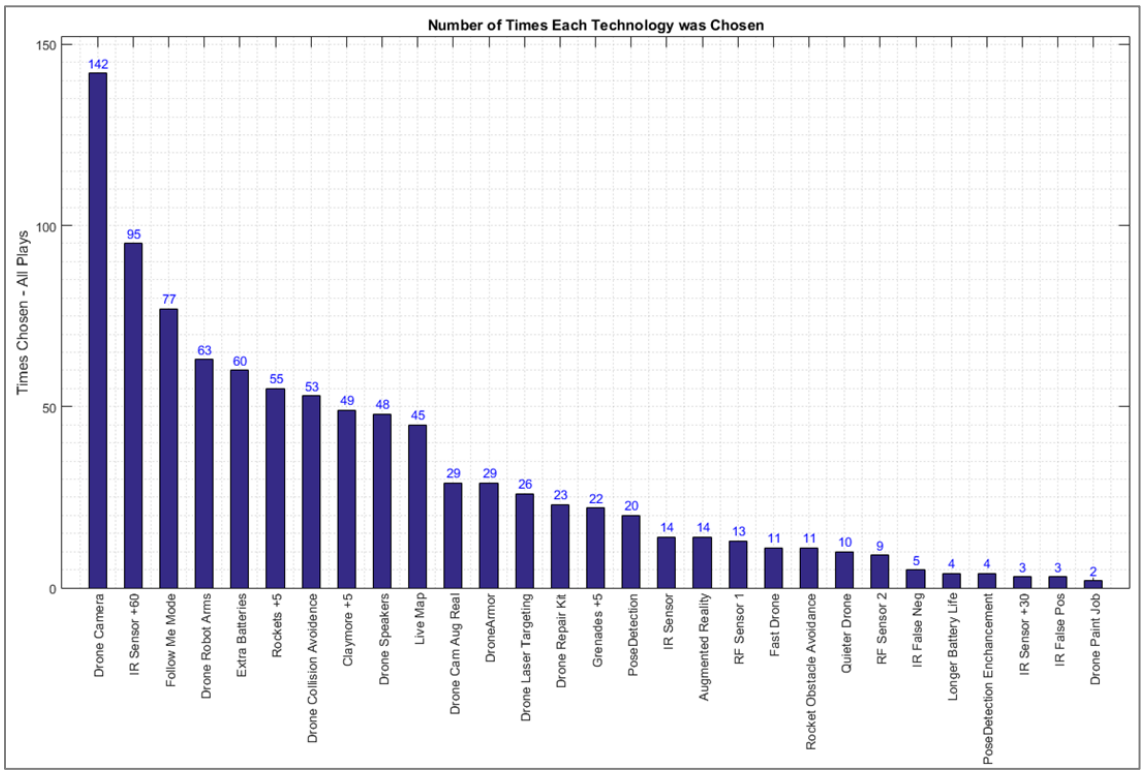

FIGURE 7. By logging the technologies players selected, the prices they were willing to pay, and the combinations they often brought together, we can get a data-driven picture of the relative utility of the proposed technologies for the modeled mission. In this case, the mission was to use a personal drone aircraft to find a lost asset inside a hostile urban environment.

Assessing players' preferences only makes sense if one believes that players are making good choices for themselves. To allay that concern, we can look for correlations in the data between players' preferences and their performance; such a correlation is shown in Figure 8. This relationship in the data helps to validate two important assumptions: (1) in-game scoring motivates players to succeed, and (2) players are honestly expressing their opinions in the technology selection mechanism. We verified the first assumption by demonstrating that players change their level of risk aversion when the score penalty for coming under enemy fire is adjusted. Even with no real-world prize at stake, players who were given higher penalties for being shot within the game showed greater risk aversion in their behaviors and technology selections. We validated the second assumption by using technology selection mechanisms drawn from economic and mathematical game theory. We used methods that are known to encourage players to be honest in their assessments of value and to not incentivize gaming the system or lowballing a bid.

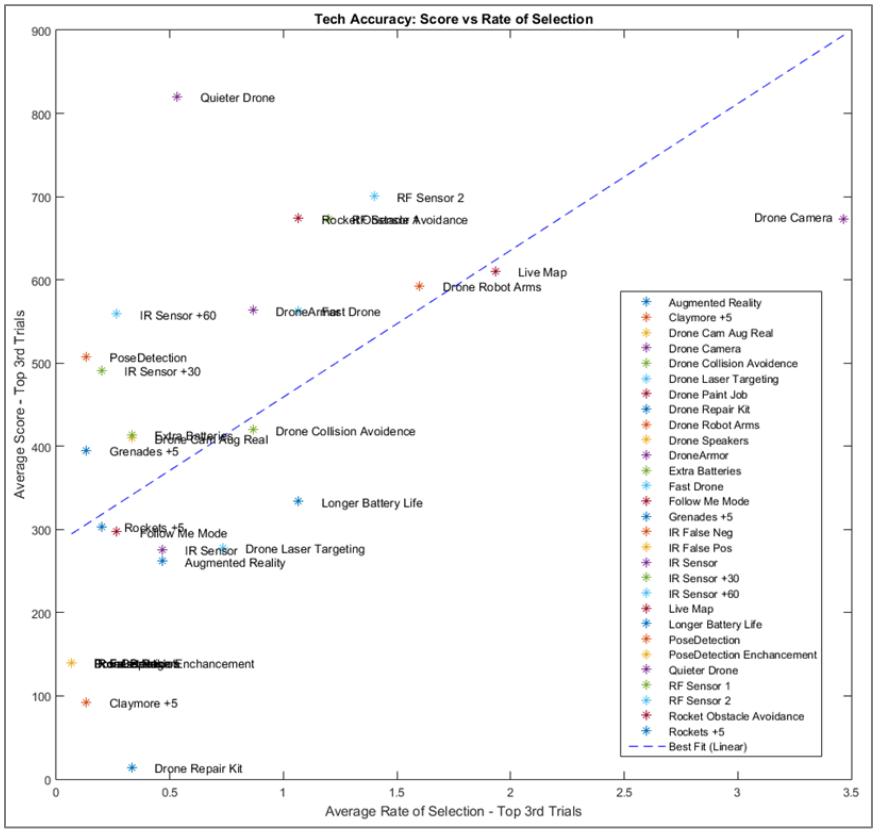

FIGURE 8. By using selection methods grounded in economic game theory and auction theory, we increase our confidence that the preferences players express for different technologies reflect their true valuation of those technologies. We see a correlation between the technologies players preferred to select and the technologies that produced better in-game performance scores.

At this point, we have reason to believe that players are forming opinions in the time provided, that those opinions reflect actual utility within the game, and that the game reflects realistic levels of risk aversion. So, we can trust the assessments players made of the modeled technologies, at least within the bounds of the mission they performed, the quality level used to model the technologies, and the correct calibration of the scoring incentives. As seen earlier, the strategies discovered by players sometimes surprise the research team, meaning that the method is capable of providing novel insights into how the technology will alter current practice.

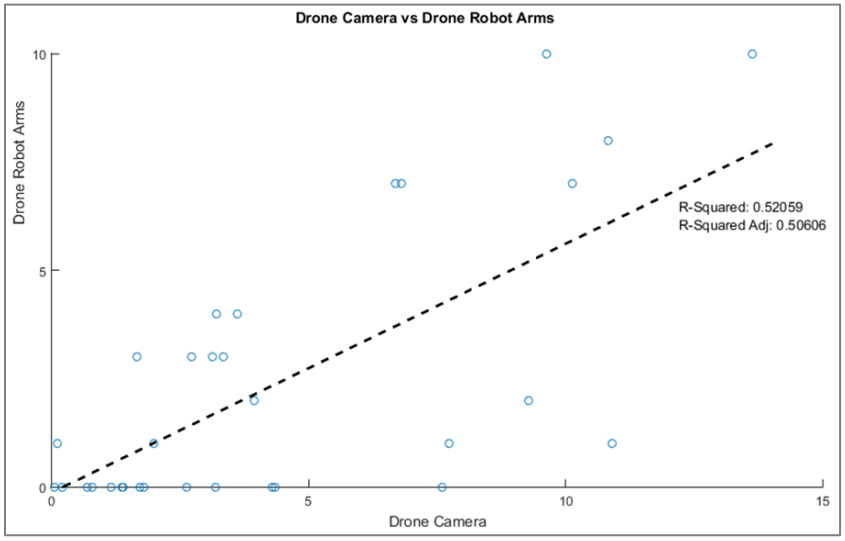

FIGURE 9. We can see here that the drone camera and drone robot arms are correlated in their selection, suggesting that they are more valuable together than individually. By looking at the synergies expressed in the data, we can identify effective suites of technology as well as individually valuable technologies. These results are statistically weaker than the individual capability assessments because of the sample size used, but they indicate a promising possibility for what we can learn from data collected from rapid-play games.

Assessing the individual value of technologies is one thing, but part of the challenge of early-phase RDTE is looking at effective technology suites, that is, combinations of technologies or capabilities that will enhance performance. So, what we'd really like to discover is which technologies are synergistic, providing more value than the sum of their parts when deployed in concert. Figure 9 shows how data collected from rapid-play games might be used to answer that question by providing correlations in the selection of certain pairs of capabilities. In the illustrated example, there is a correlation between the use of drone-mounted cameras and drone-mounted manipulative arms, indicating that each of those technologies is more valuable when paired with the other. In contrast, technologies such as IR and RF sensors show no correlation—the value of each of those sensors is independent of whether or not the other is available.

Moving Forward

The broad field of serious games is growing but still early its maturity. By and large, it has been established that digital games can be an effective tool for training users and changing their behavior, but techniques for doing so consistently and reliably are still an ongoing are of research [5]. The HIVELET work ongoing at Lincoln Laboratory aims to address that gap by providing and validating a framework for systematically modeling a domain and collecting useful data from it. In general, Lincoln Laboratory's work on serious games focused on making games a data-driven field for supporting quantitative analysis, thereby leveraging the Laboratory's data-analysis and domain-analysis strengths. Our view tends to be that a game is a sensor for measuring human decision making, thereby providing a quantitative way to study

and learn from human experts. Thinking of a game as a sensor helps frame how it can be applied to systematically evaluating both technology and user performance.

Much of the research on serious games focuses on education, training, and medical therapy, and deals with the question of transference, that is, whether or not skills or behaviors learned in a game will transfer to the real world. A smaller portion of the field, including much of the research ongoing at Lincoln Laboratory, is examining the use of games in broader roles, such as domain analysis, technology evaluation, or crowdsourcing. Traditional tabletop games and professional wargames do explore all of those areas [6], but they are typically not executed in a data-driven or iterative fashion. Our continuing research effort is to tackle problems traditionally targeted by qualitative methods and supplement them with quantitative assessment from rapid-play digital games.

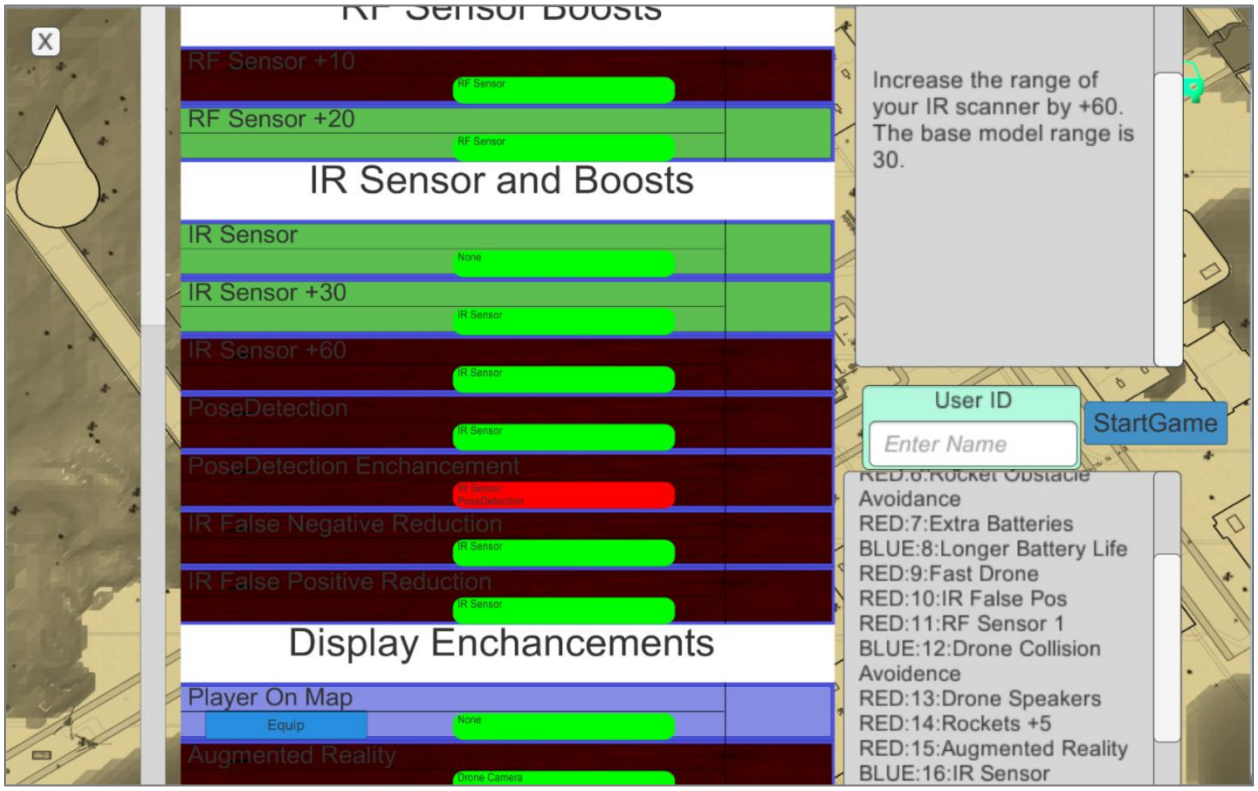

FIGURE 10. An alternate selection method drawn from game theory is a draft. There is no cost to selecting a technology, but each time the player takes a technology, the red force (either another human playing the adversary or a computer simulating an adversary) excludes three items from the list, forcing the player to prioritize their selections and avoid brittle combinations.

The HIVELET work done thus far has used a resource-constrained market as the selection mechanism that forces players to make cost-benefit assessments of proposed capabilities. A market method drives a player to find a minimalistic solution that will let them succeed at the mission. Other selection methods drawn from game theory may be effective at collecting

different types of data. For example, cake-cutting (where one player divides the set of capabilities into two groups and the other selects which group they prefer) or drafting (illustrated in Figure 10) focuses players on what combinations of technologies are most synergistic or most redundant, and a draft (where players alternately select the available capabilities) focus players on selecting flexible capabilities and building robust strategies that do not rely on any one capability being present. For different programmatic objectives, different techniques can be swapped into the framework to produce different types of data.

The mission simulator described in this article was a 3D real-time model of tactical situations. The HIVELET approach can also be paired with turn-based strategic simulators that are used to assess how capabilities might impact higher level decision making. Lincoln Laboratory has done prior work on rapid-play games for strategic level decision making, such as the one shown in Figure 11. We have not yet combined such games with the HIVELET approach; analysis of the viability of such a combination is expected in 2017.

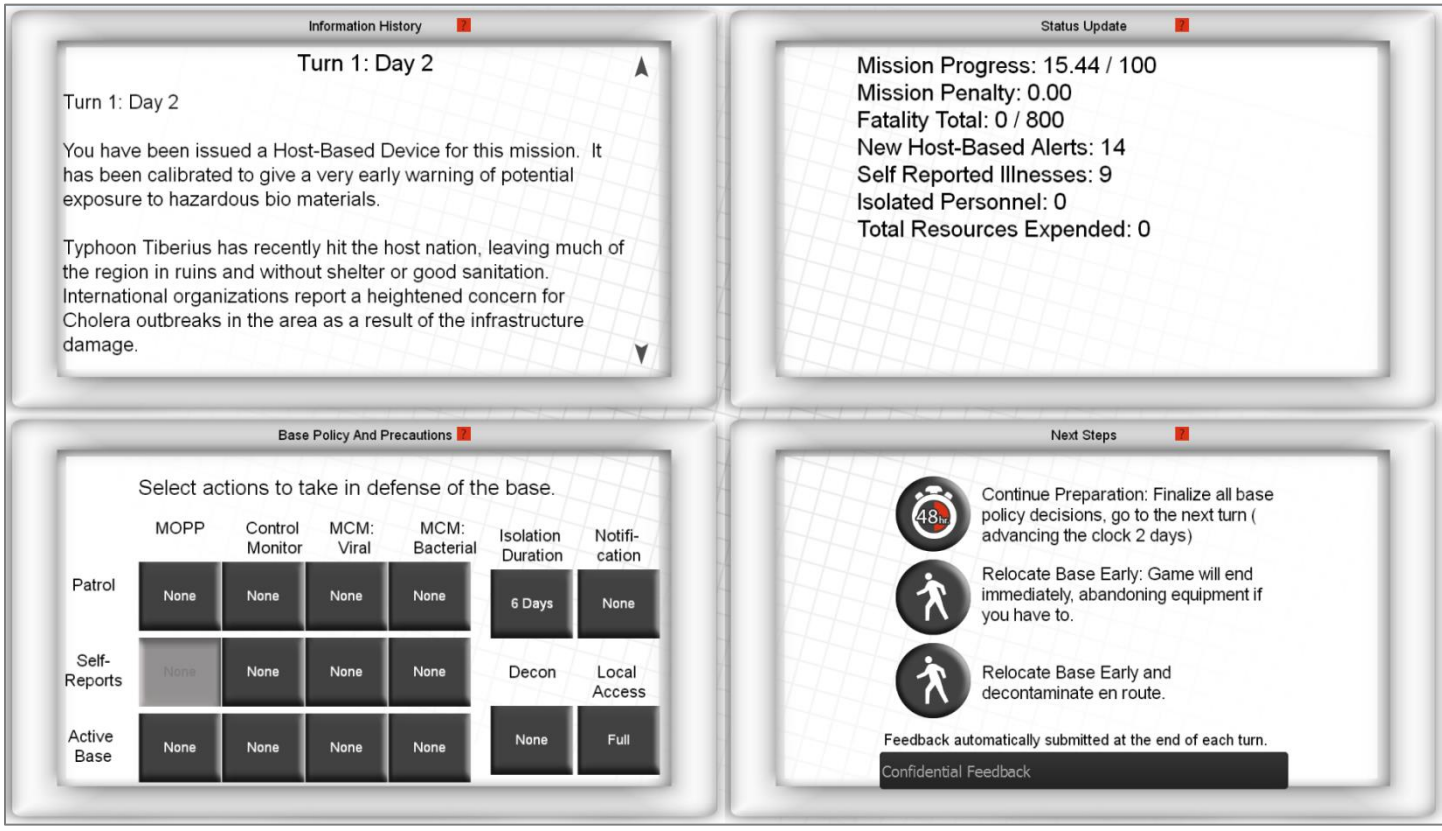

FIGURE 11. A dashboard style interface for a rapid-play game that focused on strategic-level decision making. In this game, players are managing a forward operating supply base that has potentially come under biological attack. Players use proposed future capabilities to help determine what precautions are appropriate and how much to jeopardize the mission to protect base personnel.

The infantry example described earlier in this article focused on a single-player experience facing an automated threat. Multiplayer cooperative and competitive modes need to be explored further to determine if the HIVELET technique can also provide insight into how technology changes team dynamics and adversarial situations. Multiplayer implementation of HIVELET is not a technically challenging extension, but it complicates the collection of data and thus may require many more plays before statistically meaningful conclusions can be reached. Research into the proper design of both the games and experiments will be important to broadening the work in that direction. Many emerging technologies focus on how multiple users interact, so providing quantitative support for the prioritization of technology that improves team coordination and effectiveness will be a growing field of interest that HIVELET aims to strengthen [7].

The most important piece of future work will be the application of the HIVELET technique to additional problem domains to refine, and further validate the technique so that it can be integrated more smoothly into the RDTE process.

Acknowledgments

We would like to acknowledge the contributions of the HIVELET research team—Andrew Uhmeyer, Joel Kurucar, Daniel Hannon, Joseph Isaac, and Paul DiPastina; the HIVELET supervising committee—Timothy Dasey, Thomas Reynolds, Catherine Cabrera, Paula Ward, and Martine Kalke; and Kathryn Lannin and Dorothy Ryan, who provided technical writing input.

References

1. G. Klein, *Sources of Power: How People Make Decisions*. Cambridge, Mass.: MIT Press, 1998.
2. P. Suarez, "Games for a New Climate: Experiencing the Complexities of Future Risks," The Frederick S. Pardee Center for the Study of the Longer-Range Future, Task Force Report, Boston: Boston University, Nov. 2012.
3. P. Perla, *Peter Perla's The Art of Wargaming: A Guide for Professionals and Hobbyists*, J. Curry, ed. Annapolis, Md.: U.S. Naval Institute Press, 2012.
4. A.K. Dixit and B.J. Nalebuff, *The Art of Strategy: A Game Theorist's Guide to Success in Business and Life*. New York: W.W. Norton & Co., 2011.
5. T.M. Connolly, E.A. Boyle, E. MacArthur, T. Hainey, and J.M. Boyle, "A Systematic Literature Review of Empirical Evidence on Computer Games and Serious Games," *Computers & Education*, vol. 59, no. 2, 2012, pp. 661–686.
6. D. DellaVolpe, R. Babb, N. Miller, and G. Muir, *War Gamers' Handbook: A Guide for Professional War Gamers*, S. Burns, ed. Newport, R.I.: U.S. Naval War College, 2013.
7. S.C. Sutherland, C. Harteveld, G. Smith, J. Schwartz, and C. Talgar, "Exploring Digital Games as a Research and Educational Platform for Replicating Experiments," *Proceedings of the 2015 Northeast Decision Sciences Conference*, 2015.

About the Author

Robert M. Seater is a researcher in the Informatics and Decision Support Group at Lincoln Laboratory. He currently works on serious games, requirements analysis, and software engineering. He has applied serious games to a range of topics of interest to the Department of Defense and Department of Homeland Security, including the integration of unmanned aerial vehicles into infantry squads, large-scale emergency response, chemical and biological defense, and naval missile defense. He holds a bachelor's degree in mathematics and computer science from Haverford College and a doctorate from MIT in computer science and requirements engineering.

This material is based upon work supported under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.