Taylor & Francis
Taylor & Francis Group

Theoretical Issues in Ergonomics Science

ISSN: 1463-922X (Print) 1464-536X (Online) Journal homepage: https://www.tandfonline.com/loi/ttie20

Selecting and assessing challenge problems

Jared Freeman, Drew Leins, Conrad Bell IV & The SD2 Research Consortium

To cite this article: Jared Freeman, Drew Leins, Conrad Bell IV & The SD2 Research Consortium (2019) Selecting and assessing challenge problems, Theoretical Issues in Ergonomics Science, 20:1, 27-38, DOI: 10.1080/1463922X.2018.1485987

To link to this article: https://doi.org/10.1080/1463922X.2018.1485987

Published online: 12 Jan 2019.

THEORETICAL ISSUES IN ERGONOMICS SCIENCE
2019, VOL. 20, NO. 1, 27–38
https://doi.org/10.1080/1463922X.2018.1485987

Selecting and assessing challenge problems

Jared Freeman[a], Drew Leins[a], Conrad Bell IV[b] and The SD2 Research Consortium[1]

[a]Aptima, Inc., Arlington, VA, USA; [b]ECS Federal LLC, Fairfax, VA, USA

ABSTRACT

Organisations conducting research programs often focus the work of their scientists and technologists on challenge problems (CPs). These challenges are designed to ensure that progress is measurable and relevant to the goals of the program sponsor. Generating and selecting pertinent CPs is difficult, as is assessing their value. We describe a method of generating and selecting CPs and its application in a highly collaborative, multi-organisation research program. Thirty-eight biologists, chemists, mathematicians and computer scientists across academic, commercial and government organisations generated and ranked their top choices from among 12 richly described candidate challenge problems. A ranked-choice voting formula was applied. Five CPs were highly scored; the remaining seven were distributed across a lower range of scores. The program sponsor subsequently directed researchers to address six CPs, including the elected five. Analysis of the rationales that participants offered for their CP rankings revealed four domain-independent dimensions of value: capability, speed, impact and synergy. These dimensions of value can help managers of interdisciplinary research programs systematically select a portfolio of CPs that will efficiently apply utilise resources towards program goals and facilitate measurement of scientific progress.

ARTICLE HISTORY
Received 6 March 2018
Revised 20 May 2018
Accepted 4 June 2018

KEYWORDS
Challenge problem; problem definition; research policy; voting

Relevance to human factors/ergonomics theory

Doing so addresses the significant human factors challenge of managing complex research teams and tasks in a manner that returns high research value for the funder's dollar.

Introduction

> *We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organise and measure the best of our energies and skills.*
>
> John F. Kennedy, President of the United States, 12 September 1962.

CONTACT Jared Freeman freeman@aptima.com Aptima, Inc., 1010 North Glebe Rd. Suite 700, Arlington, VA 22201, USA

[1]Members of the SD2 Research Consortium are listed in the Acknowledgements section.

Challenge propels discovery and innovation. President John F. Kennedy set the bold challenge of a manned moon-landing in May of 1961. Success on scientific, technical and political fronts followed an investment of eight years and $155 billion dollars (adjusted for inflation). Today, many Federal research organisations routinely define scientific *challenge problems* (CPs) to guide research. For example, one public notice from the Intelligence Advanced Research Projects Activity (IARPA) states:

> *IARPA intends to measure the utility of proposed solutions [on the HECTOR program] by trying to solve IARPA-furnished challenge problems. An example challenge problem might be to perform various properly authorised statistical analyses on census data without compromising privacy of any individuals …*
>
> Intelligence Advanced Research Projects Activity (2017).

CPs serve at least two major functions. First, they drive investment and progress towards some distribution over basic research, use-inspired basic research (*Pasteur's quadrant*, per Stokes 1997) and applied research. Kennedy's moon-landing challenge focused primarily on achieving a single, practical mission; as a side effect, it produced fundamental advances in geology, chemistry, medicine, material science, physics and other sciences (National Aeronautics and Space Administration 2004). Another applied research effort, the Manhattan Project, had a similar effect in advancing our fundamental knowledge of physics. The program from which we report findings has a portfolio of CPs with long-term goals of creating new scientific methods and technologies and accelerating discovery in complex areas of biology and chemistry. However, each of these CPs will also advance understanding and generate solutions for a more immediate, practical problem. In sum, programs define CPs to distribute science and technology investments across multiple levels of research in a planned way.

The second function of CPs is to support measurement of progress. CPs typically define very few goals; for example, get a man to the moon. Therefore, the metrics of progress can be standardised across research participants. Defined goals are often tangible. Therefore, the metrics can be easily understood and communicated. Measurements help program managers and their institutions promote and defend funding decisions. Measurements also provide evidence of the effectiveness of research products, evidence that can persuade operational and industrial users to fund 'last mile' transition from the laboratory to the field.

If CPs drive and justify investment, then it is important to select relevant CPs and to understand the source of their value. Doing so addresses the significant human factors challenge of managing complex research teams and tasks in a manner that returns high research value for the funder's dollar. There is, however, a dearth of research that directly provides insights on how to generate, select and value scientific problems. Consequently, we turn for inspiration to literature in the closely-related fields of defining problem spaces, solving problems and discovery.

Al-Ghassani and colleagues (2006) proposed a rational process for defining a problem space. They suggested that the research begin with a loose characterisation of the problem. Specific components and processes for addressing the problem are then identified, including key theories, methods, technologies and datasets. Finally, a

determination is made regarding how these components and processes fit together. This survey of problem and solution information tends to define the problem space crisply enough that the research can, consistent with Pólya's dictum (Pólya, 1957): understand the problem, plan a route through it, travel that route and test for success, failure and discoveries. One implication for challenge problem designers is that they pilot these steps to ensure that: (1) the problem can be expressed and comprehended; (2) the solution components – theory, science and technology – exist and are discoverable and (3) any technical interfaces between the components exist or can be constructed.

McComb, Cagan, and Kotovsky (2017) characterised the space of structural engineering problems quantitatively in three ways. First, there must be alignment between all objective functions (e.g. 'achieve a design that is light and strong') and that these functions may influence the shape of the design space. Secondly, the *local* structure of the design space indicated the sensitivity of solutions to parameter values and, thus, the likely efficiency of local searches. Thirdly, the *global* structure of the design space bounded the number of potential solutions. Accordingly, CP design should include a careful examination of goals or objective functions by which solutions will be measured. Another implication is that CP design should construct a map of the problem space, one that measures the domain and, if possible, partitions it into areas that are quantitatively or qualitatively distinct.

In the rich literature concerning human problem solving, we find rationalist and naturalist threads that provide some guidance concerning CP design. Newell and Simon (1972) typify the rationalist approach. They proposed that problem solving is a function of the solver's choice of problem representation (or model) and of one or more strategies (or *programs*) from a catalogue of heuristic search strategies. The main implication for CP design is that there be multiple distinct and promising representations of the problem (and participants willing to use them). This will increase the variety of solution strategies and the likelihood of discovering a successful solution.

Newell and Simon noted that their information processing approach did 'not shed much light' on the matter of how humans select models and strategies. The naturalist school of problem solving research attempted to reverse engineer this black box. The theory of recognition-primed decision making (Klein 1993) asserted that domain experts should recognise the problem at hand as an instance of a case in memory, retrieve a solution and apply mental simulation to test and revise the solution. The implicit guidance for CP design is that the problem be carefully distinguished from similar but solved cases to ensure that researchers solve for the new, distinguishing features.

Cohen, Freeman, and Thompson (1998) described a process of metacognitive control of recognitional decision making and generation of novel solutions under uncertainty, high stakes and time constraints. A training method based on this theory had large, reliable and positive effects among relatively expert decision makers. The key implication for CP design derives from the framework: CPs should present sufficient uncertainty, impact and urgency to inspire deep and critical thinking.

Klein (2013) presented a theory of discovery (a phenomenon adjacent to problem solving) to account for historic cases that seem to violate the simple model by Wallas

(1926), in which the discoverer engages in preparation, incubation, illumination and verification. Klein argued that discoverers attend to surprising connections and contradictions and use these as cues to test and revise theories or mental models. This guides us to design CPs in which extant theory or methods fail to explain, predict or generate relevant data. This implication would be minor if CPs typically generated data sufficient to produce surprise and drive progress, but many CPs do not.

In summary, the literature concerning CP design is scant or non-existent, although related literature offers some guidance for CP designers. At least some solution components (theory, methods, data) should populate the problem space. It should be possible to map that space, to draw or represent it in multiple ways to ensure multiple solution opportunities. The CP should be well-distinguished from solved cases and sufficiently rich in uncertainty, impact and urgency to inspire focused and critical thought.[1] A view across this literature reveals the common requirement for domain expertise in CP design. In multi-disciplinary research programs, it is generally not the case that an individual or group from one scientific discipline can design a good CP. For example, it might be unreasonable to expect a biologist to generate a CP of significant interest and value to a computer scientist. In the program on which we report, researchers had such domain expertise. Consequently, the program manager tasked them with generating a set of CPs relevant to the program's goals. They more than obliged by generating a set of CPs far exceeding the research program's limited pool of resources (e.g. time, money and labour). This raised a significant question. By what means could the set of CPs be evaluated to identify a subset of those most worthy of investment?

Below, we describe a method for systematically selecting CPs and report an analysis that will help program managers to characterise the value of a portfolio of challenge problems.

A method for collaborative generation and selection of challenge problems

The method reported here grew out of a research program with aggressive scientific goals coupled with a high degree of collaboration. Researchers in the program consisted of biologists, chemists, mathematicians and computer scientists arrayed in more than a dozen teams. These researchers gathered together at a recent workshop, coming from 18 geographically distributed organisations of different types from both the academic and commercial sectors. The program sponsor asked workshop participants to generate CPs that engaged each other in significant discovery and invention in biology, chemistry and analytics. Below, we present the methods and findings of this effort to generate and select challenge problems, and findings from a subsequent analysis of the value that researchers placed on these CPs. This process is illustrated in Figure 1.

The researchers generated 24 candidate CPs independently and in small teams. They described CPs using a template that required: the CP name; the author and collaborators; short- and long-term goals; automation requirements for execution of experiments, data processing and analysis steps; alignment with sponsor objectives and relevant scientific literature. They were well-prepared to provide this information

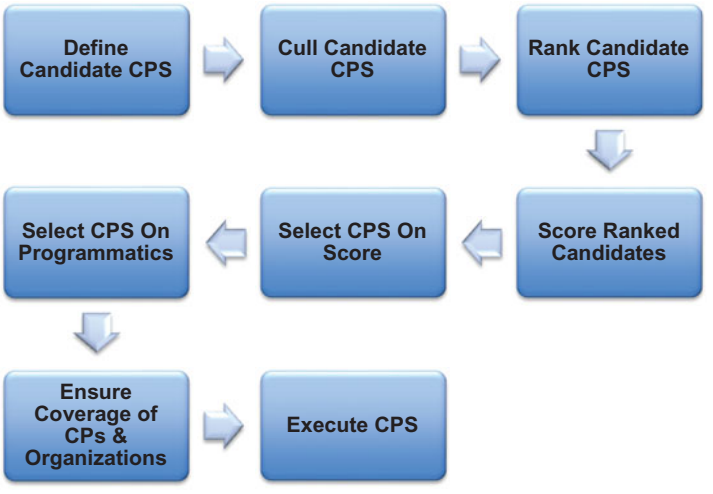

Figure 1. A process for defining, selecting and characterising the value of challenge problems.

because many participated in drafting a guiding document that laid out a range of potential applications; experimental issues; data screening and analysis issues; product components; issues of experimental reproducibility and initial concepts for challenge problems.

As an example, one researcher proposed a candidate CP to develop a detector for a dangerous chemical agent. The CP's first quarter goal was, 'Design and test [products] that detect [chemical components of threat X] with the maximum possible dynamic ranges'. The first year goal was, 'Repeat workflow on up to 100 [chemical components of threat X] in parallel'. At four years, the goal was to detect and neutralise the threat. Data from this CP, noted the author, would be processed by existing software workflows as well new analysis software being developed by another program researcher.

Researchers independently read most or all of the CPs. In a large meeting, they collectively culled 12 CPs that were redundant of others or that were subsets of other CPs.

Participants were then asked to rank order the five CPs they most preferred from among the remaining 12 CPs. Thirty-six researchers completed this task, although some ranked fewer than five CPs. We performed and informal analysis to identify five highly ranked CPs. Problems that were most frequently among the top five selections were prioritised above all others. Conflicts in rankings (ties) were resolved by selecting CPs that were most frequently among the top two choices. The top five problems were reviewed by the sponsor, who recommended extending the initial list to include a sixth highly ranked CP that closely aligned with program objectives. This informal method was fast, but not necessarily reproducible.

A formal method was subsequently applied to verify the rankings before the sponsor and research team invested heavily in the selected CPs. This algorithm, reported in Appendix A, accounted for the incompleteness of the rankings. This incompleteness arose because each participant ranked only their top five (or fewer, in some cases) of 12 CPs, not all 12.

Table 1. Quantitative analysis of CP rankings, ordered by a score that incorporates rank and frequency data, and compensates for rankings of a subset of all candidates by all participants.

CP ID	Formal score (rank)	Informal rank findings
h*	0 (1)	2
x*	5 (2)	4
a*	19 (3)	1
u*	36 (4)	3
c*	58 (5)	6
M	80 (6)	n.a.
K	88 (7)	n.a.
v*	93 (8)	5
b	97 (9)	n.a.
j	98 (10)	n.a.
d	106 (11)	n.a.
g	107 (12)	n.a.

CPs marked with "*" were selected for execution.

Shortly thereafter, a survey was conducted to elicit information about the level of engagement in the six selected challenge problems. This survey was presented as a table of CPs by performers. Each participant populated cells of their row with a score to indicate whether they: '(1) definitely can work the CP now; (2) probably can; (3) cannot determine yet; (4) probably cannot; (5) definitely cannot'. The survey results were used to ensure that each researcher could engage in at least one CP, and to identify any CP that was without a research team. The results were also used by the sponsor as a resource management tool to avoid potential resource conflicts or shortages. For example, a researcher might be directed to engage in a particular challenge problem as a subject matter expert or because a required capability was housed at their institution. Alternatively, a specific combination of researchers might be directed to engage because they were all required to successfully complete a challenge. Finally, the program sponsor would direct certain researchers away from one or more challenges to avoid the possibility of resource shortages, as would be the case if researchers distributed their resources too thinly across too many CPs.

Table 1 lists, for each CP, (1) the formally computed score and (2) the rank generated from summary ranking statistics. The analysis largely confirmed the informal findings with the exception that one CP (ID = v) was not among the five best ranked, but was roughly equivalent in its middling rankings to five others. In addition, the sponsor identified one CP (ID = c) to pursue that was not among those identified informally but that later emerged in formal analysis as the fifth highest ranked CP. The survey of engagement opportunities revealed a reasonable distribution of scientists across CPs. The program manager made modest adjustments to their self-assignments based on these data and other information.

Analysis of the dimensions of CP value

An analysis was conducted to identify the dimensions of value on which workshop participants ranked the candidate CPs.

Table 2. Four dimensions of CP value rarely co-occurred.

	Capability	Speed	Impact	Synergy
Capability	1	–	–	–
Speed	−0.12755	1	–	–
Impact	−0.32329	−0.04495	1	–
Synergy	−0.15177	0.035406	0.035806	1

Thus, they were weakly correlated and relatively independent of one another.

Method

In addition to ranking CPs, the scientists submitted rationales for their selection of each CP they ranked (in the task, above). These rationales varied in depth. A striking example of a compelling rationale was the following:

This project has a good chance of yielding a high impact publication by broadly advancing the state-of-the-art in [component] design… There are good opportunities here to combine biophysical models and [machine learning] to improve [component] function predictions, which can broadly utilise [other teams'] capabilities. Once the data tables are created for this effort (could be a Q2 milestone), the [computer scientists] will be able to dive deep with [machine learning] to identify patterns of interest.

Two of the authors independently coded a random sample of 20 of the 169 rationales that the 36 program members submitted with their CP rankings.[2] These authors then negotiated four codes that expressed most of the dimensions each had independently observed in the data, and jointly applied those codes to the same 20 rationales. One researcher coded the remaining rationales. Each CP rationale was coded as belonging to from zero to four dimensions.[3]

Results

We identified four dimensions of value[4] in CPs:

- Capability = The CP creates a tool or technique that enables new methods of analysing, designing, or producing parts, components or systems (50 instances in CP ranking rationales).
- Speed = The CP enables the program to demonstrate success quickly (22 instances).
- Impact = Discovery, design, technology, techniques and applications arising from this CP will be significant (e.g. publishable, award-worthy) (56 instances).
- Synergy = The CP builds on extant products of the program, resources of the performing organisations, commercial products, or open sources tools and data (52 instances).

These four dimensions were independent of one another. An analysis (see Table 2) over the 112 rationales in which at least one dimension was present found only weak correlations ($r < |0.33|$) between the dimensions. Thus, coders who use these dimensions in the future can do so with some confidence that they will not conflate the dimensions one for another.

Discussion

A procedure for generating and down-selecting CPs was developed and applied. It was sufficient to enable program scientists to specify CPs, to eliminate redundant or low-value CPs and to down-select among CPs using rank-based voting. A survey of participant capability to engage in each CP was also conducted to confirm that no researchers or CPs were 'orphaned'. The method drove investment decisions by the sponsor and engagement of the program team. The method did not fully capture scientific or programmatic concerns that led the sponsor to amend the results with one additional CP, as noted above. Thus, we advise users of this method to ensure systematic review and revision of CP selections by program leadership.

An analysis was conducted on the rationales that the scientists offered for their CP rankings. That analysis revealed four independent dimensions of CP value: capability, speed, impact and synergy. These dimensions have useful applications. First, a program sponsor might build a portfolio of CPs whose distribution over dimensions matches their investment strategy. For example, a portfolio whose CPs maximise speed and synergy would execute quickly and potentially at low cost. To build a 'biggest bang for the buck' portfolio, the sponsor would select CPs that maximise impact and synergy. An investment in scientific techniques and technology might involve CPs that score highly on capability development. In sum, different sponsors will likely want to use mixtures of CPs based on the overall program goals and funding.

Second, a program sponsor might analyse the relationships of dimensions, researchers and CPs to understand and manage the complex socio-technical system spawned by a multi-million dollar research investment. A correlational, cluster or principal components analysis of scientists by their use of CP dimensions in rationales should reveal their preferences for building new capability, publishing at speed, attaining high impact or building research networks. An analysis of researchers by CPs will shed light on their perceptions of their technical capabilities. Analysis of CPs by CP dimensions should reveal the perceived potential of each CP to develop new methods or technologies, demonstrate progress quickly, discover new scientific or technical phenomena, or build the research community.

In future research, we plan to pursue the applications and analyses, above. This will require a more fine-grained dataset, either historical or newly created. We plan to have each scientist rate each CP (e.g. from 1 to 7) on each of the four dimensions, as well as rank their preferred CPs. The correlation between value ratings and CP rankings will provide some validation of the dimensions, indicate which dimension(s) drive rankings of CPs,[5] and enable the correlational and dimensionality reductions described above.

The financial cost of CPs is, of course, a factor in calculating return on investment. This method did not directly capture such costs. It does give some indirect guidance: the survey of potential researcher engagement with CPs can help a program manager to manage redundant investments, and the synergy dimension may help the sponsor and scientists distribute tasks to those who can execute them most cost effectively. In the CP assessment exercise, above, neither the sponsor nor the scientists could assess the cost of the CPs, given the nascent state of the problem descriptions. Precise cost estimates require investment in deep design of research studies and, potentially, pilot

testing. The program at hand is developing automated cost estimation technologies (e.g. for laboratory costs) that will partially address this aspect of RoI.

Finally, we note that the CPs referenced here have, as of this writing, been selected but not executed. Once the CPs have been executed, a post-assessment will provide data to determine whether the program participants selected CPs that met their claims for timely execution, synergy, creation of new capabilities and impact.

Notes

1. A range of interesting research bears on the design of organizations for solving CPs, such as teams of researchers. Narayanamurti and Odumosu (2016) studied successful research institutions to define aspects of culture, policy, management, and investment. The National Research Council (2015) recommended that collaborative research programs apply lessons from the rich literature concerning teamwork. CPs, particularly interdisciplinary ones, should be designed with some expectation of the need for and cost of these practices.
2. Some scientists submitted fewer than the requested five ranked CPs, and some scientists omitted rationales for some rankings. Thus, the number of rationales is lower than expected from 36 individuals ranking five CPs.
3. A rationale was coded as having zero dimensions when the text could not reasonably be considered a rationale, or was blank.
4. Note that the name of a dimension indicates only that the rationale addressed the issue (e.g. impact) but not whether the rationale was positive or negative (e.g. high or low impact) on that dimension.
5. The dichotomous dimension data developed here did not support this type of analysis.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory under Contract No. FA8750-17-C-0294 (and related contracts by SD2 Publication Consortium Members). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA), the Department of Defense, or the United States Government. This article was approved for Public Release, Distribution Unlimited (Distribution Statement 'A') by DARPA on 18 May 2018. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the US government. This document does not contain technology or technical data controlled under either US International Traffic in Arms Regulation or US Export Administration Regulations.

We thank Jesse B. Freeman for developing the mathematical formalisation of the rank voting algorithm. We thank Jacob Beal and this journal's anonymous reviewers for insight and recommendations that improved this article.

The members of the SD2 Research Consortium, as of the date of manuscript submission, are: William J. Allen, Texas Advanced Computing Center; Carrie Arnold, Texas Advanced Computing Center; Andrew Avila, Pacific Northwest National Laboratory; Bryan A. Bartley, Raytheon BBN Technologies; Jacob Beal, Raytheon BBN Technologies; Diveena Becker, Ginkgo Bioworks; Conrad Bell IV, ECS Federal; Vanessa M. Biggers, Transcriptic; Amanda Brower, Massachusetts Institute of Technology; Tracy Brown, Texas Advanced Computing Center; Daniel Bryce, SIFT, LLC.; Diogo M. Camacho, Wyss Institute, Harvard University; Richard Cardone, Texas Advanced Computing Center; James P. Carson, Texas Advanced Computing Center; Daniel Cetnar, Pennsylvania State University; Emory M. Chan, Lawrence

Berkeley National Laboratory; Yang Choo, Transcriptic; Katie J. Clowers, Ginkgo Bioworks; James J. Collins, Wyss Institute, Harvard University/Massachusetts Institute of Technology; Alexander Cristofaro, Massachusetts Institute of Technology; Anastasia Deckard, Geometric Data Analytics; Douglas Densmore, Boston University; S. Charlie Dey, Texas Advanced Computing Center; Yuval Dorfan, Massachusetts Institute of Technology; Robert G. Egbert, Pacific Northwest National Laboratory; Hamed Eramian, Netrias, LLC; Mohammed Eslami, Netrias, LLC; Amin Espah Borujeni, Massachusetts Institute of Technology; Erik S. Ferlanti, Texas Advanced Computing Center; John M Fonner, Texas Advanced Computing Center; Jared Freeman, Aptima; Sorelle A. Friedler, Haverford College; Dany Fu, Boston University; Niall Gaffney, Texas Advanced Computing Center; Tomas Gedeon, Montana State University; Christopher Geib, SIFT, LLC; John Gentle, Texas Advanced Computing Center; Sarah Goldberg, University of Washington; Robert P. Goldman, SIFT, LLC; D. Benjamin Gordon, Massachusetts Institute of Technology; Steven B. Haase, Duke University; Sean M. Halper, Pennsylvania State University; Nathan O. Hodas, Pacific Northwest National Laboratory; Ayaan Hossain, Pennsylvania State University; Anagha Jamthe, Texas Advanced Computing Center; Manu Mary John, Texas Advanced Computing Center; Christopher Jordan, Texas Advanced Computing Center; Michael Keller, Texas Advanced Computing Center; Benjamin J. Keller, University of Washington; Eric Klavins, University of Washington; Ugur Kuter, SIFT, LLC; Peter L. Lee, Transcriptic; Sarah Leinicke, Boston University; Drew A. Leins, Aptima; Sahil Loomba, Wyss Institute, Harvard University; Julia Looney, Texas Advanced Computing Center; Eriberto Lopez, University of Washington; Chun-Yaung Lu, Texas Advanced Computing Center; Narendra Maheshri, Ginkgo Bioworks; Vikash K. Mansinghka, Massachusetts Institute of Technology; Paul Maschhoff, Ginkgo Bioworks; Joseph Meiring, Texas Advanced Computing Center; Benjamin N. Miles, Transcriptic; Vincent Mirian, Massachusetts Institute of Technology; Thomas Mitchell, Raytheon BBN Technologies; Francis C. Motta, Duke University; Tramy Nguyen, Raytheon BBN Technologies; Alexander J. Norquist, Haverford College; Rhys A. Ormond, Transcriptic; Michael Packard, Texas Advanced Computing Center; Ian M. Pendleton, Haverford College; Alex Plotnick, SIFT, LLC; Andrew Punnoose, Ginkgo Bioworks; Alexander C. Reis, Pennsylvania State University; Nicholas Roehner, Raytheon BBN Technologies; Howard M. Salis, Pennsylvania State University; Ulrich Schaechtle, Massachusetts Institute of Technology; Joshua Schrier, Haverford College; Jedediah M. Singer, Two Six Labs, LLC; Jawon Song, Texas Advanced Computing Center; Devin Strickland, University of Washington; Joseph Stubbs, Texas Advanced Computing Center; Xun Tang, Pennsylvania State University; Steve Terry, Texas Advanced Computing Center; Lisa Tiberio, Raytheon BBN Technologies; Virgina Trueheart, Texas Advanced Computing Center; Joshua Urrutia, Texas Advanced Computing Center; Matthew W Vaughn, Texas Advanced Computing Center; Christopher A. Voigt, Massachusetts Institute of Technology; Mark Weston, Netrias, LLC; Weijia Xu, Texas Advanced Computing Center; Yaoyu Yang, Ginkgo Bioworks; Enoch Yeung, Pacific Northwest National Laboratory; Jing Zhang, Massachusetts Institute of Technology; Gregory J. Zynda, Texas Advanced Computing Center; Gary Cattabriga, Haverford College.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Jared Freeman, Ph.D., is Chief Scientist of Aptima and a Principal Investigator on the SD2 program in which this work was developed. His research addresses problem solving, human learning and organisational assessment and design.

Drew Leins, Ph.D., is an experimental psychologist specialising in legal and cognitive psychology. He focuses on decision making in high-stakes environments and has worked with federal law enforcement and the military to conduct research on psychological issues critical to national security.

Conrad Bell IV, Now in his second career, Mr. Bell is a Ph.D. candidate in Computer Science and a Systems Engineering and Technical Advisor. In his first career of 25 years, he leveraged the enterprise-wide computer systems he designed and implemented into large-scale process re-engineering efforts that resulted in improved organisational efficiency in Fortune 50 companies.

Funding

This work was sponsored by the Department of Defense under contract FA8750-17-C-0294 (and related contracts by members of the consortium members).

References

Al-Ghassani, A. M., J. M. Kamara, C. J. Anumba, and P. M. Carrillo. 2006. "Prototype System for Knowledge Problem Definition." *Journal of Construction Engineering and Management*, 132 (5): 516–524.

Cohen, M. S., J. T. Freeman, and B. T. Thompson. 1998. "Critical Thinking Skills in Tactical Decision Making: A Model and a Training Method." In *Decision-Making Under Stress: Implications for Training and Simulation*, edited by J. Canon-Bowers and E. Salas. Washington, DC: American Psychological Association Publications.

Intelligence Advanced Research Projects Activity. 2017. *HECTOR*. Accessed 28 February 2018. https://www.iarpa.gov/index.php/working-with-iarpa/requests-for-information/hector?id =984.

Klein, G. A. 1993. "A Recognition-Primed Decision (RPD) Model of Rapid Decision Making." In *Decision Making in Action: Models and Methods*, edited by G. A. Klein, J. Orasanu, R. Calderwood, and C. E. Zsambok, 138–147. Westport, CT: Ablex Publishing.

Klein, G. 2013. *Seeing What Others Don't*. Philadelphia: Public Affairs.

McComb, C., J. Cagan, and K. Kotovsky. 2017. "Optimizing Design Teams Based on Problem Properties: Computational Team Simulations and an Applied Empirical Test." *Journal of Mechanical Design* 139 (4): 041101-01–041101-12.

Narayanamurti, V., and T. Odumosu. 2016. *Cycles of Invention and Discovery*. Cambridge, MA: Harvard University Press.

National Aeronautics and Space Administration. 2004. *NASA Facts: Benefits from Apollo: Giant Leaps in Technology*. Accessed 5 February 2018. https://www.nasa.gov/sites/default/files/80660main_ApolloFS.pdf

National Research Council. 2015. *Enhancing the Effectiveness of Team Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/19007

Newell, A. and H. A. Simon. 1972. *Human Problem Solving*. Upper Saddle River, NJ: Prentice Hall.

Pólya, G. 1957. *How to Solve It*. Garden City, NY: Doubleday.

Stokes, Donald E. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington, DC: Brookings Institution Press.

Wallas, G. 1926. *The Art of Thought*. New York, NY: Harcourt, Brace & Company.

Appendix A

The ranked-choice voting formula developed in this work computes an interval score for each of many candidate CPs from ordinal rankings by multiple rankers who may rank only a subset of candidates, may rank a unique subset of candidates, and may award some candidates the same rank (i.e., a tie).

The parameters are:

- n denotes the number of rankers,
- k denotes the number of candidates to be ranked,
- m denotes the maximum rank assigned over all candidates and rankers.

We will let C_j denote candidate j for $j = 1, \ldots, k$. The ordering of rankers and candidates is fixed and identified with the sets $\{1, \ldots, n\}$ and $\{C_1, \ldots, C_k\}$.

For each of the n rankers, we define a ranking function:

$$r_i : \{\text{Candidates}\} \to \{\text{Ranks}\} \cup \{m+1\}$$
$$\{C_1, \ldots, C_k\} \to \{1, \ldots, m, m+1\}$$

Above, $m+1$ corresponds to a candidate that a ranker does not rank. This score represents a worst case, in which the ranker considered all candidates not ranked as equal in priority and minimally less desirable than the lowest ranked of the ranked candidates. We define a pre-weight as follows:

$$\tilde{w}(C_j) = \sum_{i=1}^{n} r_i(C_j)$$

We define the weight function $w(C_j)$ using the following normalisation, in which C_i is the candidate with the lowest pre-weight:

$$w(C_j) = \tilde{w}(C_j) - \min_{i \in \{1,\ldots,k\}} \tilde{w}(C_i)$$