Modeling Analytics as Knowledge Work: Computing Meets Organizational Psychology

Guangchao Yuan[1], Nirav S. Ajmeri[1], Chris Allred[2], Pankaj R. Telang[3], Mark Wilson[2], Munindar P. Singh[1]
[1]Department of Computer Science, [2]College of Humanities and Social Science, [3]Laboratory of Analytic Sciences
North Carolina State University
Raleigh, NC, 27695-8206
Email: {gyuan, najmeri, cmallred, prtelang, mark, mpsingh}@ncsu.edu

***Abstract*—This paper reports on an ongoing interdisciplinary study of analytic workflow, describing our preliminary understanding and findings as well as some directions for further investigation and validation. Specifically, we exploit knowledge from organizational psychology to develop a computational organization model. Our proposed organizational model provides a framework to understand the impact of organizational level variables and worker characteristics on workflow performance, providing a view to create justifiable interventions to improve performance. To evaluate the viability of the model, we develop a multiagent simulation framework and design an experimental study.**

I. Introduction

Analytic workflow describes how analysts, collaborating with colleagues and supervisors, apply personal knowledge, tools, and organizational resources to perform their work. Our research objective is to understand the factors that govern analytic workflow performance to identify interventions that improve the effectiveness and efficiency of analytic performance. Analytic workflow in our terminology includes human and organizational aspects of how analysis is conducted.

The broader relevance of analytic workflow to information systems and science arises from the rapidly expanding interest in Big Data analytics, for use in industry, clinical research, and elsewhere. For this purpose, we can think of analytic workflow as (generally cooperative) knowledge work focused on answering questions. Existing research on analytics focuses on algorithms for analyzing data at scale and on visualizing Big Data. The deep challenges of how people formulate questions and hypotheses and how they arrive at concrete recommendations based on the data are not tackled in current work. We claim that these challenges are particularly relevant to understanding how information systems are used within the broader challenges of organizations that perform analytics.

Workflow in a lot of traditional computer science research, such as on business process management and cyberinfrastructure for science, is studied from a low-level perspective in which the specific tasks and constraints on the mutual ordering are given high importance. These representations focus excessively on task structure and information flow at the cost of not modeling the human aspects of workflow adequately. Frequently, such approaches face resistance from workers. They are rarely deployed outside of restricted settings such as insurance claims processing. Indeed, they are fundamentally unsuited to representing knowledge work.

We consider two complementary aspects of analytic workflow. The *operational domain* deals with micro-level workflow in which workflow is modeled at the individual analyst level, within a given context, for a given task. Work at the operational level will seek predictive models of performance based on people, tools, and their mutual relationships in fixed organizational contexts so as to produce interventions in end-user representations and tools. The operational system model emphasizes analysts (and tool-supported interactions); it yields operational performance metrics.

The *organizational domain* refers to macro-level considerations including general workflow trends, the formal and informal organizational context (e.g., work norms, reward structure, and culture). Work at the organizational domain seeks predictive models of performance based on people, organizations, and their mutual relationships so as to produce interventions such as training, culture interventions, and staffing. The organizational system model emphasizes the organization and roles responsible for key organizational metrics, and yields organizational performance metrics.

The research questions of interest fall into three groups.

- Building a quantified understanding of analytic workflows. Within this, sample questions may include (Q1) What are suitable metrics at the operational level for both inputs (e.g., tool complexity, user personality type) and performance outputs (e.g., formalizing quality, efficiency)? (Q2) What are suitable metrics at the organizational level for both inputs (e.g., nature of interactions, supervision, expertise, and affective engagement) and outputs (e.g., repeatability of processes, compliance with regulations)?
- Predicting performance in analytic workflows. (Q3) What is a predictive model of performance at the operational level based on analyst and contextual attributes (e.g., does a disparity in response times of two tools leads to biases by preferring the faster one)? (Q4) What is a predictive model of performance at the organizational level (e.g., is peer adoptions related to tool adoption across contexts)?
- Applying the understanding to produce improvements in analytic workflows. (Q5) How can we determine effects of interventions (e.g., external peer reviews reduce confirmation bias; forced repository usage lowers quality).

Contributions

In this paper, we present a formal metamodel for analytic workflow in an organization, including elements of tasks, work processes, and states and traits of social entities (individuals and organizational units). This metamodel, being iteratively refined and enriched over a period more than one year, will help capture important aspects of real-life workflow as it arises in organizations. Specific models of workflow in specific organizations built using this metamodel capture how users communicate and otherwise interact with one another as they jointly carry out their work. For example, a communication from a supervisor may alter a users emotional state and may alter the state of their team to one of increased cohesion. In addition, we present our preliminary evaluations of the organization model.

Paper Structure

The remainder of the paper is organized in the following sections. Section II provides some additional motivation for our research. Section III reviews and discusses previous study in the organizational research. Section IV presents our proposed organizational metamodel. Section V introduces our preliminary designed evaluations of the model. Section IV concludes the paper and discusses our next plan.

II. Motivation and Approach

We consider three primary approaches to investigate analytic workflow.

Work analysis. This approach involves interviews and surveys of analysts to identify realistic use cases of analytic workflow and work functions as well as contextual factors influencing analytic workflow and performance. From this exercise, we hope to understand objective functions capturing trade-offs between effort, accuracy, confidence, and resource usage, and constraints such as urgency and the risks of different types of errors.

Laboratory study. This approach employs proxy problems/analysts engaging in analytic workflow in a controlled environment. This approach allows us to refine methods of assessing analytic workflow and to investigate potential alterations to analytic workflow.

Computational modeling. The third approach involves computational modeling of the operational and organizational domains, producing a basis for predicting performance. The associate models accommodate states and traits of individuals, organizations, work products and processes, and tools.

From these approaches, we seek to derive a set of recommendations for modifications in workflow and possibly in the contextual variables to enhance performance. Our laboratory environment supports the investigation of potential recommendations while our computational models support a simulation of organization-wide ramifications of potential analytic workflow changes. The work-analytic approach can help inform the feasibility of potential improvements.

The main challenge faced by our research is that it requires fundamental understanding the impact of organizational level variables and worker characteristics on workflow and performance with a view to creating justifiable interventions to improve performance. To inform the development of our model, we conduct an interdisciplinary study with researchers from computer science and organizational psychology. Our research includes two iterated steps: (1) review previous literature and conduct discussions; (2) collect data and perform experiments.r

We propose to use agent-based organizational simulations to model the effects of organizational characteristics and individual differences on workflow and performance. Key to this approach is the creation of realistic real-time organizational information using a combination of field data collection and data fabrication techniques informed by existing research on relationships among organizational characteristics, individual differences, and performance. Although initial organizational models will be simple, the intent is to expand them based on field data to mirror the actual organization being simulated. Doing so will enable the validation of simulated outcomes with respect to actual outcomes and the exploration of organizational states and concomitant outcomes, including those that may have a low probability of occurrence in the actual organization. Importantly, because of their incorporation of the proposed data fabrication engine, our simulations would be "meta-simulations" that parametrically support families of workflows and organizations.

In addition, we present our preliminary evaluations of the organization model. Specifically, we want to investigate how do the states and traits of individuals affect performance of a workflow. We develop a multiagent simulation framework and present our design for an experimental study. The two kinds of evaluations are interdependent. The agent-based simulation informs the design of the experiment study, whereas the experiment study learns the algorithms that will used in the simulation. It should be noted that our organizational model does not restrict to the research question that we investigated in our evaluation section.

In future work, we plan to (1) collect realistic data from results established in the literature, organizations, and our designed experiment study; (2) validate the simulation with respect to data about tasks, work processes, individuals, and organizations acquired through pilot field studies; (3) build a data fabrication engine from our simulation results and experiment study; (4) explore the impact of potential changes in the organization on workflow and performance that are driven by the data fabrication engine; (5) cross-validate the effectiveness of the agent-based organizational dashboard against actual organizational workflows and outcomes; (6) enhance the dashboard so that it learns from its the success or failure of its predictions.

III. Background

Organizations are complex systems and characteristics of organizations such as individual and team working, formation, and workflow interests researchers from various fields. These characteristics have been studied extensively but independently in psychology, management, engineering and computer science. Ours is an interdisciplinary study that brings in researchers from different communities to better understand the analytic workflow in an organization from multiple perspectives.

Further, as the researchers are inevitably outside of the organization in which the target users work, several organizational characteristics and behaviors remain inaccessible to the researchers and thus remain unstudied. Agent-based modeling and simulation provides a way to study these characteristics and behaviors [1], however it is underused in organizational psychology [2]. We evaluate our model through both agent-based simulation and experiment study.

Our work relates to Crowder et al.'s framework [3] for simulating engineering team work. Crowder et al. [3] propose an agent-based modeling approach for simulating working of a team in an engineering environment, based on research in two engineering organizations. Their model includes variables at individual level, team level and task level, and incorporates *DesignerAgents* who perform assigned tasks, *ResourceAgents* who respond to information seeking requests, and *TaskManagerAgent* who allocates tasks to *DesignerAgents*. Their simulation model identifies how team performance change if these variables are changed.

Hsu [4] proposes a complexity-based approach for intra-organizational team selection. Hsu's study employs a computation model and uses a small design firm as a case to compare the performance of different team selection approaches such as random selection and equity method, and replacement policies in different economic conditions. Hsu simulates the team-selection approaches using an agent-based model. The simulation results show that the interdependence-based selection team selection methods can yield better performing teams than traditional ability-based team selection methods. It also suggests that managers should protect higher-performing workers, and transfer low performers to another team before replacing them. Our computational model emphasizes on how and what kind of factors would govern the performance of analytic workflows. Even though our work focuses on understanding the workflow on an analyst, it can be adopted to study and evaluate team work in general.

IV. Organization Model

We produced an organization model after multiple discussions among researchers from organizational psychology and computer science, and interviewing experts in the organizations we are studying. Our developed organization model is shown in Figure 1.

Social Entity. Social entities include both ORGANIZATIONS and INDIVIDUALS, where an individual is a part of an organization. In addition, an organization is hierarchical: an organization may belong to another organization.

States. States are attributes of a social entity that are transient or at least have the flavor of being transient. Classic examples include the emotional state of a user. Satisfaction and organizational commitment are important in organizational research because they are related to withdrawal behaviors (e.g., turnover) [5]. States in our model represent an individual's satisfaction and commitment toward tasks.

Traits. Traits represent an individual's attributes that are associated with some stability. Classic examples include personality and cognitive ability. Even though a user may become more competent over time, we might attributed her success in terms of increasing training and experience playing on stable traits such as innate language ability.

Role. An organization defines a role for an individual that requires certain QUALIFICATIONS, imposes certain LIABILITIES upon an adopter of a role, and granting PRIVILEGES to the individual. Meanwhile, the role restricts the position an individual may actually adopt.

Position. A position is the job that an individual actually performs in an organization. For example, "faculty" is the role that an university defines, but every faculty member may not do the same job even though their roles are the same. We differentiate "position" from "role" because we hypothesize that the misalignment between position and role will affect a workflow's performance.

Work Process. A work process is the scheduling of TASKS, representing an overall task that could be measured by performance. Tasks can be completed sequentially, in a parallel or overlapping manner, and may depend on the completion of preceding tasks [6].

Tool. An individual would have skills to apply certain tools to complete a task.

Performance. Performance can be categorized into *process* and *objective* performance. Process measures capture how a work process performs in terms of EFFICIENCY and EFFECTIVENESS, while objective measures how a work process in terms of QUALITY and QUANTITY.

Social Resource and Preference. Each individual has "social preferences" for the team workers that she wants to collaborate with, and has "social resources" that she could employed to perform a task. The "social resource and preference" of a work process could be difficulty and required duration of tasks, the order of tasks, and so on.

V. Evaluation

The model enables us to study a series of research questions to understand the factors that govern analytic workflow performance. To evaluate the viability of our model, we start with a research question: how do the states and traits of individuals affect performance of a work process? We have highlighted the modules that will be involved in investigating the question in Figure 1.

We evaluate our model through agent-based simulation and experiment study. The two kinds of methods are interdependent. The agent-based simulation informs the design of the experiment study, whereas the experiment study learns the algorithms that will used in the simulation.

A. Simulation

We implement an agent-based simulation using JADE [7]. The simulation framework enables us to analyze the effect of each variable in a work process's performance efficiently, offering a view to create justifiable interventions to improve performance.

Agents in the Simulation: The simulation model has four types of agents: an organization agent, a manager agent, many task agents, and many customer agents. The organization agent represents the organization; the manager agent schedules tasks within a work process and assigns tasks to tasks agents; the task agent is the one that actually performs tasks. A

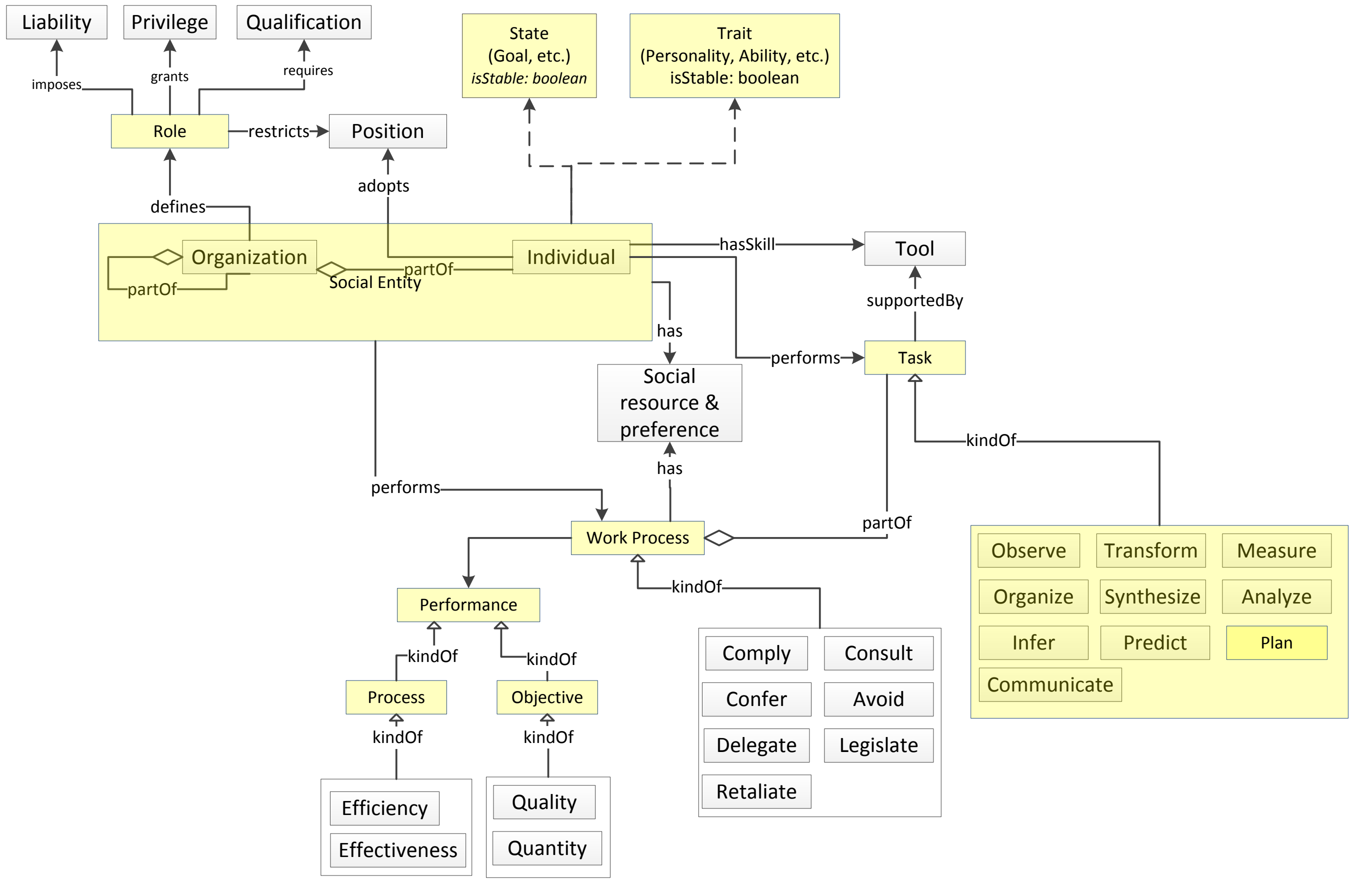

Fig. 1. Organizational model.

customer agent submits its requirements for the work process to the manager agent, and the manger agent may accept or reject the requirements. If the manager agent accepts the requirements, it will initiate several work processes that satisfy the requirements to execute. Table I summarizes the behaviors of different agents types in our simulation framework.

TABLE I. SUMMARY OF BEHAVIORS OF AGENT TYPES IN THE SIMULATION FRAMEWORK.

Agent type	Behavior
Organization agent	Assigns role and position to each task agent Always responds to resource seeking requirements Announces state and trait of a task agent
Manager agent	Negotiate with customer agents Assigns tasks to task agents Computes performance of a work process
Task agent	Performs assigned tasks Seeks required resources from the organization agent
Customer agent	Submits requirements to the manager agent

Modeling a work process: We model a work process by dividing it into a number of tasks, where each task is undertaken by a single task agent. Tasks can be undertaken in parallel or sequentially. A work process may be subject to constraints such as that some tasks should to be undertaken sequentially; that is, a task can only start if its preceding tasks have been completed. Such constraints are specified before the simulation starts. Each task has a task type, some required skills, and a deadline. The task will only be assigned to a task agent if the agent has the required skill.

Algorithms: Algorithms in the simulation framework capture how different variables correlate with each other, and how these would variables affect the performance metrics. We obtain algorithms in two ways: by analyzing data collected from the experiment study and via discussions with experts in organizational research. For the analyses of data collected from the experiment study, we conduct multiple regression analyses to obtain the numeric solutions between the variables. Specifically, the algorithms in our simulation framework include the following three functions.

State transition table. We assume that each task agent's state will be updated after executing a task, and the organization's state will be updated after a certain period.

State-Trait-Task-Quality (STTQ) table. We assume that the quality of each task depends on the state and trait of each task agent, and the attributes of the task. The attributes of a task include the task type, the required duration, and so on.

Actual time of completing a task. We assume that the actual time for each task agent to complete a task depends on capability of each agent and the quality of each task. That is, for the same task, the agent that has higher capability requires less time to complete the task; for the same agent to perform the same task, it requires more time to make the resulting quality higher.

Simulation procedures: We have implemented the simulation framework and can run a series of extensive simulations given the input of work processes. The simulation runs in the following steps:

1) A customer proposes a requirement
2) The manager accepts or rejects the requirement
3) If the manager accepts the requirement, he maps the requirement to several work processes, and assigns tasks to workers
4) Each work accepts the assigned task, and chooses a task from his queue to execute
5) The manager computes the performance once a work process is done
6) Each task agent updates its state after executing a task

B. Experiment Study

We conduct experiment study to learn features of the algorithms used in our simulation model and understand in abstract terms how the attributes of the model affect the outcomes.

A major challenge to a study such as the present is to obtain the correct parameter settings in an evidence-based manner. We are designing empirical studies and will conduct them shortly as a way to obtain the data to ground the simulation in real-life settings. That is, the analyses of the data will be incorporated into the simulation. We now introduce the design of our experiment study.

Participants:

- *Option 1:* The sample will consist of undergraduate students enrolled in an introductory to psychology course at a large, urban, southeastern university. Average SAT Verbal and Quantitative scores of entering Freshmen typically fall between 1175 and 1250. While demographic data will be collected, the sample is expected to be predominantly Caucasian with significant African American and Asian subgroups. The expected range for age is between 18 and 22 years of age. Participants will receive course credit for their participation.
- *Option 2:* The sample will consist of individuals over the age of 18, who are English speakers, and who have access to a spreadsheet application. Participants will be recruited and compensated monetarily using Amazon Mechanical Turk (MTurk). The amount of payment will be related to the degree of complexity of the task being completed. A posting will allow MTurk members to confidentially sign up using their randomly assigned MTurk ID. As data collection will be conducted via MTurk, no identifying information will be collected.
- *Option 3:* The sample(s) will consist of individuals described in both *Option 1* and *Option 2.*

Work Process: Each participant will be presented with a single work process, requiring the successful completion of multiple tasks in order to achieve successful performance. The work process, presented on a computer in the form of a task "worksheet" (e.g., a Word document), will consist of an analytic problem (this will vary in complexity between-participants) that will require the cleaning and manipulation of data files in the task folder.

Procedure: Informed consent will be obtained from all participants prior to the beginning of the study. Once informed consent is obtained, demographic data will be collected anonymously via Qualtrics online survey software. The experiment will be conducted during single sessions such that each participant will only be required to participate at a single time and place. Once the participant has submitted the demographic information, they will be instructed to complete an online survey designed to measure their personality as well as their capability. Once completed, they will be instructed to open a Word document (henceforth referred to as the task worksheet) located on the desktop of their personal computer containing the experimental research prompt followed by the data analytic task. Each task worksheet will contain a unique identifier that will not be traceable back to the participant. Once the prompt has been read, participants will be instructed to access the study task folder also located on the desktop. Inside the folder will be an excel files with multiple sheets that participants must access and manipulate the data within in order to successfully complete the analytic task presented on the task worksheet. Participants in the deadline-presence experimental group will be noted via the experimenter when limited time remains for each task. Following the answering of each task, participants will (1) rate their perceived task difficulty via Qualtrics as well as (2) their level of task engagement. Upon completion of all tasks, participants will save their work process worksheet and be debriefed regarding the nature of the experiment.

Research design and manipulations: A between-participants experimental design will be utilized to test the proposed research questions. Participants will be randomly assigned to an experimental group varying along two dimensions: (1) the presence of a task deadline versus the absence of a task deadline and (2) the amount of required tasks presented to the participant (4 8).

We now describe the measures we will adopt.

Personality: Personality will be assessed using the Mini-IPIP, a 20-item scale developed by Donnellan et al. [8]. The Mini-IPIP was originally developed and validated across five studies establishing the instrument as both a psychometrically acceptable and practically useful measure of the Big Five factors of personality (e.g., extraversion, agreeableness, conscientiousness, emotional stability, and intellect/imagination). This measure uses a 5-point rating scale ranging from 1 (*Very Inaccurate*) to 5 (*Very Accurate*). Respondents are asked to describe themselves in an honest manner, in relation to other people they know who are roughly the same age. Coefficient alpha will be calculated following data collection to assess the internal consistency of the scale in the study sample.

Task engagement: Task engagement will be assessed using modified versions of participant engagement items first developed by Meade and Craig [9]. The task engagement scale to be used in the current study consists of two factors; diligence and interest. Diligence will be assessed via 5-item subscale consisting of the strongest (i.e., highest factor loadings) diligence items reported.

Interest will be assessed via a 3-item subscale consisting of the strongest interest items. This measure uses a 5-point rating scale ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Coefficient alpha will be calculated following data collection to assess the internal consistency of the scale in the study sample.

Task difficulty: Perceived task difficulty will be assessed using a self-reported single item measure. SRSI Difficulty will be assessed as the response to the item: "How difficult did you find the most recent task?" This measure uses a 5-point rating scale ranging from 1 (*Very Difficult*) to 5 (*Very Easy*).

Individual capability: Participant capability will be assessed via the following items. Using a self-reported single item measure, SRSI Capability will be assessed as the response to the item: "How many college-level STEM courses e.g., Science, Technology, Engineering, Math have you completed?" Secondly, participants will be asked to input their current university GPA (for Intro to Psychology participant pool). Participants will also be asked to rate their level of use with spreadsheet applications on a 5-point rating scale ranging from 1 (*Never Use*) to 5 (*Frequently Use*). Lastly, participants will be asked to list any professional certifications that they currently possess.

Task duration: For each task, the amount of time that the participant expends effort towards solving the task will be recorded by the experimenter.

Task performance: Task performance will be assessed via a dichotomously scored item (correct, incorrect) for each sub-task response.

VI. Conclusions

In this paper, we have presented our ongoing interdisciplinary study. This study is focused on two main research questions:

- How can we model analytic workflow in a systematic manner?
- What kind of factors would govern analytic workflow performance?

We exploit knowledge from organizational psychology to develop a computational model of organizations. (Our model can is better thought of as a metamodel since it can be used to create specific models for organizations.) Our proposed organizational model provides a framework to understand to impact of organizational level variables and worker characteristics on workflow performance, providing a view to create justifiable interventions to improve performance.

To evaluate the viability of the model, we develop a multiagent simulation framework and design an experimental study. In near-term future work, we plan to (1) use our self-made analytic problems to conduct the experiment study, and run extensive simulations with the learned algorithms from the experiment study; and (2) collect real workflow data from organizations to conduct the two kinds of evaluations. The evaluations provide a way to study the effect of each factor in the workflow performance and identify interventions that improve mission enablement.

Acknowledgment

Thanks to the US Department of Defense for financial support of this research through the NCSU Laboratory for Analytic Science.

References

[1] J. R. Harrison, Z. Lin, G. R. Carroll, and K. M. Carley, "Simulation modeling in organizational and management research," *Academy of Management Review*, vol. 32, no. 4, pp. 1229–1245, 2007.

[2] H. P. N. Hughes, C. W. Clegg, M. A. Robinson, and R. M. Crowder, "Agent-based modelling and simulation: The potential contribution to organizational psychology," *Journal of Occupational and Organizational Psychology*, vol. 85, no. 3, pp. 487–502, 2012. [Online]. Available: http://dx.doi.org/10.1111/j.2044-8325.2012.02053.x

[3] R. Crowder, M. Robinson, H. Hughes, and Y.-W. Sim, "The development of an agent-based modeling framework for simulating engineering team work," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 42, no. 6, pp. 1425–1439, Nov 2012.

[4] S.-C. Hsu, "A complexity-based approach to intra-organizational team selection," 2013.

[5] K. A. Hanisch, C. L. Hulin, and S. T. Seitz, "Mathematical/computational modeling of organizational withdrawal processes: Benefits, methods, and results," *Research in personnel and human resources management*, vol. 14, pp. 91–142, 1996.

[6] Q. W. Fleming and J. M. Koppelman, "Integrated project development teams: another fad or a permanent change," *International Journal of Project Management*, vol. 14, no. 3, pp. 163–168, 1996.

[7] F. L. Bellifemine, G. Caire, and D. Greenwood, *Developing Multi-Agent Systems with JADE (Wiley Series in Agent Technology)*. John Wiley & Sons, 2007.

[8] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, "The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality." *Psychological assessment*, vol. 18, no. 2, p. 192, 2006.

[9] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data." *Psychological methods*, vol. 17, no. 3, p. 437, 2012.