Extramural Research Report 2019-01

The Validation of a Domain-General Systems Thinking Assessment Test for Personnel Selection and Classification

Tyler H. Shaw
Reeshad S. Dalal
Stephen J. Zaccaro
William Miller
Dean Cisler
Samantha Dubrow
Amanda Harwood
MaryJo Kolze
Wenmo Kong
Sam Monfort
Jake Quartuccio
George Mason University

This extramural research report is documentation of work performed under Cooperative Agreement between the U.S. Army Research Institute for the Behavioral and Social Sciences and George Mason University (Department of the Army Cooperative Agreement # W911NF-15-2-0064). This report has not received or passed a formal peer review by the U.S. Army Research Institute for the Behavioral and Social Sciences but has been submitted to the Defense Technical Information Center (DTIC) in accordance with regulatory requirements. The opinions, findings, and recommendations expressed herein are those of the authors and do not reflect official policy or position of the U.S. Army Research Institute for the Behavioral and Social Sciences or the Department of the Army.

March 2019

Approved for public release; distribution is unlimited.

Extramural Research Report 2019-01

NOTICES

DISTRIBUTION: This Extramural Research Report has been submitted to the Defense Technical Information Center (DTIC). Address correspondence to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-ZXM, 6000 6th Street (Building 1464 / Mail Stop: 5610), Fort Belvoir, VA 22060-5610.

FINAL DISPOSITION: Do not return this report to the U.S. Army Research Institute for the Behavioral and Social Sciences.

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 5 MAR 2019	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 11 SEP 2015 – 10 SEP 2016
4. TITLE AND SUBTITLE The Validation of a Domain-General Systems Thinking Assessment Test for Personnel Selection and Classification		**5a. CONTRACT NUMBER** **5b. COOPERATIVE AGREEMENT NUMBER** W911NF-15-2-0064 **5c. PROGRAM ELEMENT NUMBER** 622785
6. AUTHOR(S) Tyler H. Shaw, Reeshad S. Dalal, Stephen J. Zaccaro, William Miller, Dean Cisler, Samantha Dubrow, Amanda Harwood, MaryJo Kolze, Wenmo Kong, Sam Monfort, Jake Quartuccio		**5d. PROJECT NUMBER** 790 **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) George Mason University Psychology Department 4400 University Drive, MSN 3F5 Fairfax, VA 22030-4422		**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street (Bldg. 1464/Mail Stop 5610) Fort Belvoir, VA 22060-5610		**10. SPONSOR/MONITOR'S ACRONYM(S)** ARI **11. SPONSOR/MONITOR'S REPORT NUMBER(S)** Extramural Research Report 2019-01

12. DISTRIBUTION/AVAILABILITY STATEMENT:
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
ARI Research POC: Dr. Kristophor Canali, Selection and Assignment Research Unit
This extramural research report is documentation of work performed under Cooperative Agreement between the U.S. Army Research Institute for the Behavioral and Social Sciences and George Mason University (Department of the Army Cooperative Agreement # W911NF-15-2-0064). This report has not received or passed a formal peer review by the U.S. Army Research Institute for the Behavioral and Social Sciences but has been submitted to the Defense Technical Information Center (DTIC) in accordance with regulatory requirements. The opinions, findings, and recommendations expressed herein are those of the authors and do not reflect official policy or position of the U.S. Army Research Institute for the Behavioral and Social Sciences or the Department of the Army.

14. ABSTRACT
The goal of this research was to develop a domain independent measure of systems thinking (ST) and to collect preliminary construct and content validation evidence for the model. An extensive literature search identified four sub constructs of systems thinking: holistic thinking, adaptability, forecasting, and closed-loop thinking. Measures of these four ST constructs were developed and refined by modifying a version of the Air Force Multi-Attribute Task Battery (MATB). The criterion-related validity of two existing measures of ST (an ability/skill measure and a dispositional tendency/preference measure) via job performance ratings was assessed. A small-sample validation study showed that the four ST subconstruct operationalizations demonstrated adequate variability, were often moderately positively correlated with each other, and were related to several conceptually relevant skill/ability and dispositional variables. Findings from a larger-sample criterion-related study using previously existing measures of ST are also discussed. Directions for future research and additional refinements to the existing ST model are discussed.

15. SUBJECT TERMS
Systems theory, systems thinking, holistic thinking, adaptability, forecasting, closed-loop thinking

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
				64	Dr. Tonia Heffner
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER**
Unclassified	Unclassified	Unclassified	Unlimited Unclassified		(703) 545-4408

THE VALIDATION OF A DOMAIN-GENERAL SYSTEMS THINKING ASSESSMENT TEST FOR PERSONNEL SELECTION AND CLASSIFICATION

EXECUTIVE SUMMARY

Research Requirement:

A "system" could be a biological organism or a profit-making organization, an ecology or an employee, a religious faith or a job, a computer network or a professional network. "Systems thinking" (ST) is the tendency and ability to think about entities *as systems*—and to view systems as wholes whose components display complex causal patterns of interdependence.

Extant ST research is prolific but is greatly fragmented across a variety of intellectual disciplines and is frequently conceptual/speculative rather than empirical. Moreover, extant empirical work is frequently constrained to a particular intellectual discipline or even a particular system within an intellectual discipline; in other words, it is domain-specific. Furthermore, this work often involves self-report measures of ST, which may be susceptible to socially desirable (vs. honest) responding and measure a dispositional tendency or preference toward ST rather than an ability or skill associated with ST.

Procedure:

We reviewed the extant research literature on ST and, on this basis, proposed a more inclusive four-component conceptualization for ST. Then, consistent with practical needs associated with employee selection and (re-)classification (in the Army and civilian organizations), we developed, pilot-tested, and provided preliminary validity information for a domain-general, behavioral (i.e., skill-based) individual differences measure of ST. We selected the Air Force Multi-Attribute Task Battery (AF-MATB), itself modified from an original task developed by NASA, as our simulation of interest because of the inherent domain-general nature of the task and the ease with which the tasks could be modified to be interconnected. The MATB was originally developed to examine cognitive load and multitasking ability, though it has been used in general decision-making and systems research. We modified the AF-MATB to create the MATB-Systems Thinking (MATB-ST), which generated scores on four components of ST: holistic thinking, closed loop thinking, forecasting, and adaptive/flexible thinking.

Separately, we also examined the criterion-related validity of two existing measures of ST (an ability/skill measure and a dispositional tendency/preference measure) vis-à-vis job performance, using a larger-sample survey.

Findings:

Preliminary, small-sample validation work on the MATB-ST yielded promising results: scores on the operationalizations of the four components of ST exhibited adequate variability (thereby ameliorating concerns that the task was too easy or difficult), the operationalizations

were often moderately positively correlated with each other (but not so strongly as to be redundant), and the operationalizations were related to several conceptually related skill/ability and dispositional tendency/preference variables.

Findings from the larger-sample criterion-related study using previously existing measures of ST yielded the following findings: (a) The dispositional tendency/preference measure of ST mediated the impact of ability/skill and dispositional tendency/preference antecedent variables on job performance; (b) The ability/skill measure of ST did not exhibit appreciable criterion-related validity, pointing to the need for better ability/skill measures of ST (such as, perhaps, the MATB-ST); and, (c) Interestingly, and although we call for additional research on this topic, ST did not predict job performance better on more complex jobs than on simple jobs, despite the fact that one might expect complex jobs to require the most "systems skills."

Future Directions:

Future research should: (a) refine the MATB-ST operationalizations of ST; (b) further explicate the conceptual nomological network for ST at multiple levels of analysis (e.g., individual employee level, team level, organization level); (c) examine convergent validity, criterion-related validity, fakability, trainability, and adverse impact of the MATB-ST operationalizations of ST compared to self-report operationalizations and other skill-based operationalizations; (d) further examine the impact of job complexity on the criterion-related validity of ST; (e) examine validity in both non-Army and Army settings; and (f) construct a shorter version of the MATB-ST that can be used for mass testing in an employee selection context.

THE VALIDATION OF A DOMAIN-GENERAL SYSTEMS THINKING ASSESSMENT TEST FOR PERSONNEL SELECTION AND CLASSIFICATION

CONTENTS

Page

INTRODUCTION/PROBLEM STATEMENT 1
Problem Statement 1
Scope of Effort 2

LITERATURE REVIEW 3
Literature Review on Conceptual Mode 3
Differences Between the Original and Revised Models 7
Literature Review on Relevant Measures 7
Survey Development and Administration to Narrow the List of Systems Thinking (ST) Constructs 9

MATB-ST ADAPTATION AND DEVELOPMENT 12
Description of Air Force Multi-Attribute Task Battery (AF-MATB) 12
Modification of the AF-MATB to the MATB-ST 13
MATB-ST Performance Metrics of Systems Thinking 14

MATB-ST EXPERIMENTATION 15
MATB-ST Content Validation Procedure 15
ST Criterion Validation Research Using an MTurk Sample 25

GENERAL DISCUSSION 29
Limitations 29
Future Research 30
Implications for the US Army 32
Conclusion 33

REFERENCES 34

APPENDICES

APPENDIX A. COMPARISON WITH OTHER MEASURES OF SYSTEMS THINKING A-1

APPENDIX B. CONSTRUCT DEFINITIONS AND EXPERIMENTAL HYPOTHESES B-1

APPENDIX C. PILOT TEST RESULTS C-1

APPENDIX D. CRITERION VALIDATION STUDY TABLES AND FIGURES D-1

LIST OF TABLES

TABLE 1. FINAL LIST OF COGNITIVE AND DISPOSITIONAL MEASURES USED IN THE SURVEY 11

TABLE 2. MATB-ST STUDY EDUCATION 16

TABLE 3. MATB-ST STUDY VOCATION 16

TABLE 4. MATB-ST STUDY RACIAL IDENTIFICATION GROUP 17

TABLE 5. MATB-ST STUDY NATIONALITY 17

TABLE 6. DESCRIPTIVE STATISTICS FOR SURVEY DATA 18

TABLE 7. INTERCORRELATION MATRIX FOR RAW PERFORMANCE VALUES 19

TABLE 8. DESCRIPTIVE STATISTICS FOR THE RAW PERFORMANCE ON THE TASK 21

TABLE 9. DESCRIPTIVE STATISTICS FOR THE DISPLAYED SCORES 21

TABLE 10. COMPONENT DESCRIPTIVE STATISTICS 22

TABLE 11. THE RELATION OF 14 SCALES/ABILITY METRICS WITH OUR FOUR SYSTEMS THINKING (ST) COMPONENTS 23

TABLE 12. ITEMS USED FOR MEASURING MODERATOR AND OUTCOME VARIABLES 27

LIST OF FIGURES

FIGURE 1. ORIGINAL MODEL OF THE PROPOSED RELATIONSHIP BETWEEN THE THREE SYSTEMS THINKING (ST) VARIABLES 5

FIGURE 2. REVISED MODEL OF SYSTEMS THINKING (ST) AND PROPOSED RELATIONSHIP AMONG THE FOUR VARIABLES 7

FIGURE 3. SCREENSHOT OF THE MATB-ST. IN THIS SCREENSHOT 13

FIGURE 4. FULL MODEL FOR THE CRITERION VALIDATION RESEARCH 26

The Validation of a Domain-General Systems Thinking Assessment Test for Personnel Selection and classification

Introduction/Problem Statement

It is appropriate to conceptualize a system as "an interacting combination, at any level of complexity, of people, materials, tools, machines, software, facilities, and procedures designed to work together for some common purpose" (Chapanis, 1996, p. 22). Examples of weapons systems include missiles, tanks, and bombers. Loosely defined, our current understanding of systems thinking (ST) is that it is "an approach that views systems as wholes rather than compilations of individual components and allows one to see the interconnectedness and interdependencies of agents within systems, to frame problems as patterns, and to get at underlying causality" (Davis, Dent, & Wharff, 2015, p. 335). Systems thinking was developed as a conceptual framework more than 25 years ago (Senge, 1990), and has been applied broadly across many fields. The need to assess ST is becoming increasingly important, primarily because the systems in which military and civilian personnel are immersed are becoming much more complex. The need for ST is best illustrated by its relevance to multiple fields and disciplines, such as health care (Adam & de Savigny, 2012), social policy (Sterman, 2002), and information systems (Checkland, 1997). Even though the ST construct is represented across many fields, all representations of the construct converge upon one primary theme: the need to adapt human thought to the complexity of the world around us. Thus, it is necessary to shift from linear reductionist approaches to non-linear dynamic approaches that can address the multifaceted and interconnected relationships among several components.

Peter Senge's (1990) best-selling management book popularized the concept of ST. Because this book was written for a popular audience, however, researchers and practitioners are more inclined to work on ST that emphasizes construct validity and scientific rigor. Much academic work has certainly been conducted, but the work is greatly fragmented and in need of organization and review. Moreover, when discussing the putative outcomes of ST, both academic and practitioner sources have predominantly made a conceptual case rather than an empirical one. What is necessary for advancement in understanding ST, and for establishing the basis for selection, classification, and training of personnel in this capacity, is a comprehensive and integrated framework that provides a basis for effective measurement development and specification of its antecedents and outcomes. The present work is an initial step in that direction.

The current research report details the efforts made by George Mason University and the U. S. Army Research Institute to operationally define ST and make it amenable to measurement. The document will begin by discussing the specific goals of this effort. Next, we describe our specific approach to addressing those goals, followed by a description of our experimental methods and results. Finally, we discuss how this work is relevant to the Army and make suggestions as to how the work can be extrapolated.

Problem Statement

Given the recognized importance of ST, researchers have attempted to create interventions to enhance complex systems training. One approach to the measurement of ST has been to develop paper- or computer-based assessments. For example, Sweeney and Sterman (2000) created a series of tests that described a system and asked participants to draw a graph that forecasted the change in system behavior over time. Some of these assessments have become extremely well known in the ST community, such as the "bathtub" problem, which requires participants to graph the rise and fall of the water level in a bathtub with an open drain as the flow of water increases and decreases.

Several criticisms have been raised regarding measurement of the ST construct. Some authors suggest that these measures have not extended beyond cursory, anecdotal, and conceptual conjecture as opposed to submitting these ideas and concepts to empirical evaluations (cf. Cavaleri & Sterman, 1997). Cavaleri and Sterman (1997) also criticize the apparent disconnect between laboratory-based interventions and the way in which they can be applied to organizational performance. Unfortunately, as the aforementioned section suggests, not many empirical efforts have been made to combat these criticisms.

Many evaluations of ST have assessed cognitive change by asking participants to review their experience and describe how the intervention has altered their thinking (e.g., Cavaleri & Sterman, 1997). However, there are several issues with the validity of this type of retrospective self-report of mental events. Nisbett and Wilson (1977), for example, report the results of several experiments in which people were unable to accurately report on the factors that affected their cognitive processes. This is corroborated by general criticism of self-report that suggests that there is always some question as to whether any form of self-report accurately reflects respondents' true perceptual experiences of task performance (Natsoulas, 1967). More recent critiques are offered by Moroney, Biers, and Eggemeier (1995), who claim that self-report can interfere with task performance. Further, asking participants their opinion about the effectiveness of interventions involves providing them with detailed information about the purpose and hypotheses behind the project, which can lead to operator bias (Moroney et al., 1995).

Scope of Effort

Two principles guided our approach in developing and testing our measure of ST. The first was to adopt a skill-based approach rather than the knowledge-based approaches used in previous work. To that end, we sought to identify cognitive, metacognitive, and dispositional tendency/preference constructs that have practical and theoretical relevance in their ability to estimate ST. The second guiding principle was a focus on process-level assessment that is not tied to a particular domain. To achieve this goal, we attempted to select cognitive, metacognitive, and dispositional constructs that we thought would be associated with our ST variables. These constructs were used for content validation in our simulation. Our vision is that the constructs in our simulation will constitute our ST assessment test and that the simulation will represent a microworld (a rich computer-based simulation of a work or decision-making environment) that represents our conceptualization of the ST process (our conceptualization will be described in the "revised model" section of the Literature Review section of this document).

First, we conducted an extensive literature search on ST to develop our conceptual framework. Next, we thoroughly examined the literature for existing cognitive and dispositional constructs that were practically and theoretically related to the components of our conceptual model. The next step was to identify and develop a microworld that we could adapt as our measure of ST. Finally, we conducted content (and face) validity (Lawshe, 1975; Mosier, 1947) and convergent and discriminant validity (Campbell & Fiske, 1959) studies involving our cognitive and metacognitive predictors. We also ran a supplementary Mechanical Turk (MTurk) study to further explore our conceptual model and ST as a predictor of job performance. Specific steps made towards each of these will be detailed in subsequent sections of this report. It should be noted that throughout the duration of the research, we consulted with a subject matter expert (SME) at several stages of the project to ensure the face validity of our Systems Thinking Assessment Test (STAT).

Literature Review

One of the first tasks of the research was to conduct an extensive literature review, which involved two efforts. The first effort was to further develop our operational definition of ST. This effort required a literature review at two phases: upon the initial construction of the proposal and a "deep dive" into the ST literature upon starting the project. The literature review conducted at the beginning of the project required a much more in-depth foray into the literature. Using PsycINFO, the search term "systems thinking" brought up hundreds of articles, and we limited our search to articles with "systems thinking" in the title. Definitions were taken from articles in the first nine pages of the search. Next, two separate raters went through the 32 separate definitions that were obtained, listed key concepts from each one, and kept track of how many times each key concept was repeated. When ratings were completed, the raters came together to reach a consensus on the most important and the most often repeated key concepts that were related to ST. After this literature review iteration, we concluded that our original model required revision.

The next effort was to find existing cognitive, metacognitive, and dispositional tendency measures for ST. Importantly, the focus was on finding measures that were representative of our conceptual model. A nomological net of 47 variables that could be included was cast. Using methods advanced by Anderson and Gerbing (1991), the set was reduced to a list of 25 variables that the researchers felt were highly relevant to ST. The list was further reduced to 13, using input received from the SME in conjunction with those variables deemed most relevant by the research team. Both efforts will be described in more detail below.

Literature Review on Conceptual Model

Original model. From our preliminary investigation into this area, we believed that ST could be characterized, and thus measured, through three core components. Our preliminary literature review suggested that most of the definitions of ST cohered around three primary themes: (a) planning/learning a system, (2b) strategic management of the system, and (c) an awareness of the emergent properties of the system that are driven by changes in the

environment, as shown in Figure 1. What follows is a detailed description of our original conceptualization of the three components.

Planning/orientation style. This variable was included because what often emerged in the ST literature was a reference to the learning orientation of operators. Here, what is emphasized is an approach to training that is called learner control, in which trainees have the opportunity to select the method, timing, practice, or feedback, on all aspects of training (Milheim & Martin, 1991). It is proposed that operators who focus more on task "mastery," that is, becoming familiar with the information processing elements of performing a task, tend to have superior performance on a transfer task that requires these same process elements. Thus, what is emphasized here is the similarity of the information processing between the training and transfer tasks (Bransford & Franks, 1976). This view can be contrasted with older conceptions of task transfer, such as identical elements theory, that assert that transfer from one task to another is greater if more elements (task characteristics) are shared between two tasks (Woodworth & Thorndike, 1901). By ensuring that the information processing elements are shared between training tasks and transfer tasks, one can thus have a domain-free estimator of transfer task performance.

Strategic management. Often referred to in the ST literature as "operational thinking" or "closed-loop thinking," strategic management of the system is of critical importance to the management of complex systems. For example, Richmond (1994) viewed this as incorporating three ST skills: system-as-cause thinking, closed-loop thinking, and operational thinking. System-as-cause thinking is the notion that the structure of a system can be viewed as the underlying cause of what is driving operator behavior rather than the behavior being driven by external factors. Closed-loop thinking, then, raises the question that if the structure is the cause of behavior, what does the structure look like and how is it represented? The answer to this question, according to Richmond (1994), is that causal relations do not run one way but are instead reciprocal, in that patterns of system behavior feed back into changes in performance of the system. Operational thinking posits a structure for the way in which loops are composed. Taken together, the three thinking skills can be viewed as a way of (a) examining system structure to determine how operators will behave; (b) understanding the nature of how feedback will drive and alter behavior; and (c) having an understanding as to how the structure and the closed-loop system will constrain the feedback loops. Thus, strategic management is an understanding of the causal-loop relation and how system feedback drives performance and is primarily concerned with strategy maintenance.

Anticipation of emergent properties. Also embedded in many conceptualizations of ST is the need to predict system change and adapt accordingly. For example, Checkland (1997) has noted that understanding "the adaptive whole" is the central tenet of ST. Checkland (1997) points out, for example, that psychologists concerned with human performance often examine individual differences that affect reactions to automation failures. One problem with this approach, however, is that it is necessary to wait until a failure occurs to understand which operators cannot adequately rectify the system failure. Recently, the field has moved towards finding other ways to predict which operators are likely to be "out of the loop"—a process that makes operators more susceptible to performance decrement when system changes occur. To this end, work by Bahner, Haüper, and Manzey (2008) has indicated that you can successfully predict

failure detection by examining self-efficacy and performance indices. For example, “sampling behavior,” or the degree to which operators verify the appropriate functioning of system parameters even when the system is fully automated, has been linked to increased failure correction rate. In that study, the authors were able to simulate this activity by using a complex microworld (Sauer, Wastell, & Hockey, 2000) that allowed for all system parameters to be automated. Sampling behavior of system parameters was thus used as an index of “in the loop” thinking. It should be noted, however, that the emergent properties variable of ST is not limited to the occurrence of automation failures; it can be also related to changes in environmental structure that require a previously adopted strategy to be abandoned in favor of a new one. Thus, this variable is primarily concerned with strategy adaptation.

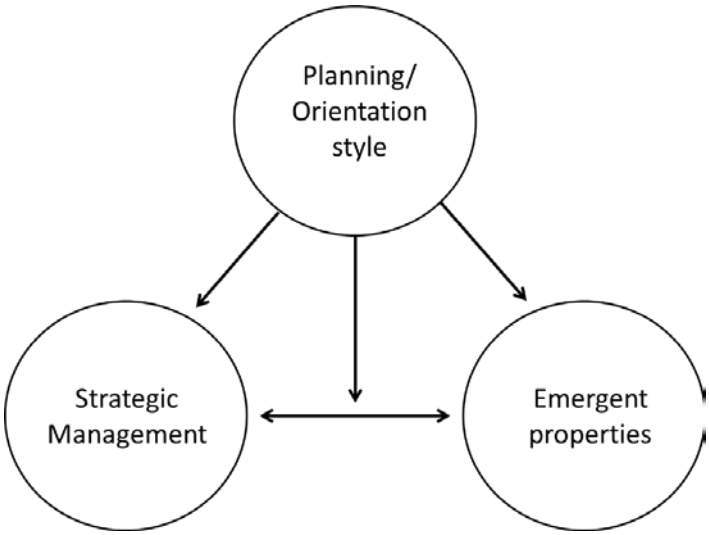

Figure 1. Original model of the proposed relationship between the three systems thinking (ST) variables.

Revised model. Upon initiation of the project, the first goal was to revisit our original conceptualization with a more in-depth literature review. To that end, we sharpened the search criteria and did a “deep dive” into the ST literature. The reason for this was simply to ensure that we had a construct that (a) synthesized the bulk of the literature as precisely and concisely as possible, and (b) was able to define the construct in such a way that made it amenable to measurement. Because of this in-depth literature review, our model changed slightly. Specifically, due to internal discussions and verbiage that had been used in the ST literature, we renamed most of our original variables and divided one component into two. What we arrived at was a four-component model of ST. The first step was to thoroughly examine the literature to uncover the way ST was usually defined. The second step was geared towards extracting what we thought were the major themes from the definitions. Finally, independent raters tallied the frequency with which each of the variables in the four-component construct was mentioned in definitions of ST. From our in-depth literature review, our new model of ST now has four components (a) holistic thinking tendency, (b) closed-loop thinking, (c) forecasting, and (d) adaptability. What follows is a description of the four components of our revised model, which is shown in Figure 2.

Holistic thinking tendency. Originally considered the planning/orientation variable in the old model, the holistic thinking tendency variable broadens that notion by not only characterizing this construct as planning and learning orientation, but also as an understanding of how overall system status is affected by its individual elements. For example, Olszewski (2014) refers to ST as "a holistic attitude, seeing interrelationships rather than seeing components separately."

Closed-loop thinking. Originally referred to as the "strategic planning" variable, closed-loop thinking has many of the elements of the original variable. Again, Richmond (1994) suggests that this incorporates three ST skills: system-as-cause thinking, closed-loop thinking, and operational thinking. System-as-cause thinking is the notion that the structure of a system can be viewed as the underlying cause of what is driving operator behavior rather than the behavior being driven by external factors. Closed-loop thinking, then, raises the question that if the structure is the cause of behavior, what does the structure look like and how is it represented? The answer to this question, according to Richmond (1994), is that causal relations do not run one way but are instead reciprocal, in that patterns of system behavior feed back into changes in performance of the system. Hence, this variable characterizes a person's understanding of the direct causal and cyclical relationships among subsystems.

Forecasting. The forecasting variable is the first of two offshoots of what was originally considered to be the emergent properties variable. The components of the original variable that can be retained here are those that refer to "in the loop" and "out of the loop" processes. As described earlier, the "out of the loop" phenomenon is a process that makes operators more susceptible to performance decrement when system changes occur. It has been thought that "sampling behavior," or the degree to which operators verify the appropriate functioning of system parameters even when the system is fully automated, has been linked to increased failure detection rate. Sampling the behavior of system parameters can thus be used as an index of "in the loop" thinking. Forecasting, then, is a time-relevant variable that refers to how system change and unanticipated occurrences are handled by the operator.

Adaptability. The second offshoot from the emergent properties variable from the original model is referred to as adaptive thinking. This would be what Checkland (1997) refers to as "the adaptive whole" component of ST. Many would define this as that variable that is concerned with action or change. The reason it is distinct from forecasting is that it requires an operator to change or alter a strategy based on feedback, whereas forecasting would have more to do with how the operator's current strategy can be projected into the future.

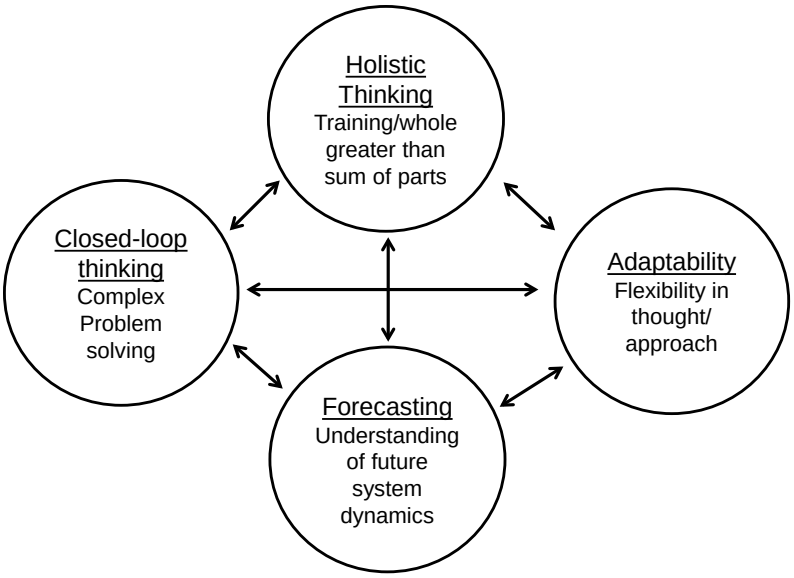

Figure 2. Revised model of systems thinking (ST) and proposed relationship among the four variables. Double-headed arrows are intended merely to indicate an expectation that the four constructs are interrelated. They are not intended as indicators of causal directionality (or bi-directionality).

Differences Between the Original and Revised Models

Two changes are easily apparent from the first to the current model. One, the model now consists of four components instead of three. Second, the interrelationships that we had proposed are now left open-ended, or bi-directional. It should be stated that there is nothing that says that all the components in the model are orthogonal—indeed, many of our variables will overlap and have shared variance. However, we think that (a) there will be sufficient unique variance contributed to the construct of ST as a whole, and (b) it was necessary to be consistent with the current nomenclature that persists in the ST literature.

Literature Review on Relevant Measures

The second effort of our literature review was to select what we originally called cognitive and metacognitive measures that were to be included in STAT. We have since changed our characterization of these measures to cognitive and dispositional measures. These measures were selected on the basis of assessments of cognitive processes that we determined were linked to our conceptualization of ST ability. Selection of these variables was driven by their theoretical relevance and empirical history.

In the selection of the initial set of variables to be included, we followed two approaches. The first approach was inspired by Cronbach and Meehl's (1955) seminal text in construct validation. In that approach, the goal is to first cast a set of measurable variables that will constitute the theoretical framework for what should be measured and the methodological

framework for how it will be measured. This constituted our nomological network. Development of a nomological network is constrained by a few key considerations. The first goal of the formation of a nomological network is to make clear what something is or means. Next, the majority of the measures and constructs in your network have to be measurable, or observable, and predict the latent constructs. Finally, operations that are qualitatively different or not related to the construct of interest should be eliminated from the definition of the construct variable. Consistent with this approach, the 11 team members were tasked with coming up with a set of measures theoretically related to the constructs in ST. We chose this approach because this would give us sufficient variability in measure selection, which we could later exploit to pare down the list into a workable set. The next section describes our approach to paring down the list of predictor variables.

Filtering for the most relevant measures. We sought to use existing methods to reduce the set of measures in our network to those deemed the most relevant. A team of 11 experts (i.e., the principal investigators and graduate students working on the project) were presented with these key concepts and agreed on a four-pronged definition of ST. The experts then compiled a list of other constructs that they believed would be related to one or more of the four aspects of ST, or ST as a whole. Thus, a set of predictor variables that was hypothesized to load on our four-factor model of ST was derived. The purpose of this step was to make the construct amenable to measurement.

The experts were instructed that while making their selections, they should think as broadly as possible, with the restriction that a brief definition and short justification for why the component should be included should be provided. To that end, 47 different cognitive and dispositional variables were included.

Clearly, 47 predictor variables are too many to be used for research purposes. Hence, the next step was to pare down the list using existing scientific methods. To that end, we followed the recommended advice from Anderson and Gerbing (1991), whose method is defined formally as a pretest methodology for predicting the performance of measures in a confirmatory factor analysis. Pretesting is an “activity related to the development of the questionnaire or measurement instrument to be used in a survey or an experiment.” It has been used as a means for reducing ambiguity in the meaning of measures. The reason for using this guide was to ensure that the measures we have selected for estimating our ST construct will tap the intended factor and not tap unintended constructs in the set. This procedure is especially recommended for field research because time and cost considerations prohibit continued access to samples of subjects large enough to permit meaningful empirical assessment of construct validity. Anderson and Gerbing (1991) support these claims by showing compiled results from two pretest samples of 20 respondents, which showed that use of their two established coefficient values could adequately discriminate measures that would be retained in a subsequent confirmatory factor analysis from those that would not. Thus, the authors have developed a pretest methodology that assesses the substantive validity of individual measures or constructs.

Substantive validity, as defined by Anderson and Gerbing (1991), is the degree to which a measure is judged to be reflective of, or theoretically linked to, some construct of interest. Assessment of substantive validity rests on judgments that are made by (a) experts, and (b)

individuals (judges) considered representative of a population of interest. To that end, the authors developed an "item-sort" task in which respondents were given a set of constructs defined in everyday language and were asked to assign each item to the one component (or overall ST) that in their judgment the item best described. Substantive validity is then assessed by two indices:

1. ***Substantive-validity coefficient (C_{sv})***: an index that reflects the extent to which respondents assign an item to its posited construct more than to any other construct.
   Formula: $C_{sv} = (N_c - N_a)/N$

   Nc = number of respondents assigning a measure to its posited construct
   Na = the highest number of assignments of the item to any other construct in the set
   N = total number of respondents

Values range from −1 to 1, with larger values indicating greater substantive validity. Large, negative values for C_{sv} also would indicate that an item has substantive validity, but for a construct other than the one posited by the researcher.

2. ***Proportion of substantive agreement (P_{sa})***: the proportion of respondents who assign an item to its intended construct.
   Formula: $P_{sa} = N_c/N$

   N_c = number of respondents assigning a measure to its posited construct
   N = total number of respondents

Values range from 0 to 1, with larger values indicating greater substantive validity.

Survey Development and Administration to Narrow the List of Systems Thinking (ST) Constructs

In developing the survey, certain characteristics needed to be taken into account. The first was how to specifically classify how each component would map onto the variable of interest. To that end, a classification scheme was developed in which all the variables could be classified in one of the following ways:

- Antecedent: An event preceding or occasioning another event; that is setting the stage for a particular response;
- Component: A constituent element, as of a system;
- Outcome: An end result; a consequence;
- Not closely related: We do not expect even a small correlation between this and ST, and
- Multiple: Can load on multiple components.

The first survey administration consisted of an *n* of 11, whereas the second consisted of an *n* of 18. The survey was administered online and took about 45 minutes to complete.

Initial findings to narrow the list of ST constructs. We applied the aforementioned procedure to our set of constructs. We took a two-stage approach (consistent with what was mentioned above) to dwindle down the list. The first stage was to have the original team of 11

researchers (defined by the principle investigators and students working on the project) take the survey that was constructed (in an item-sort fashion). Variables from the earlier literature review that were found to not have been closely related to the ST variable were considered to be a candidate for elimination. The experts judged other variables to be either an antecedent or a component of ST, which qualified those variables as candidates for retention. These are based on the criterion of a .5 for the proportion of substantive agreement coefficient.

Based on the results of the first stage, the goal of the second stage was to further pare down the list by having individuals (judges) who were considered representative of the population of interest follow the same procedure as before. Because our target population was college-aged individuals 18 to 23 years of age, we constructed a new survey based on the outcomes of the initial stage and distributed it to George Mason University graduate students. Once this stage was completed, we were in a position to assess which predictor variables should be retained and which should be excluded. This stage helped us further pare down the list and arrive at a manageable set.

After using the procedures above, we identified a preliminary set of candidate measures for our final results from the expert ratings and the graduate students as shown in Table 1 below. The initial set of candidate measures can be viewed in Table 1. We refined the list further to include only those measures that we thought were most relevant to ST. In addition, we compiled a set of instruments that can capture each measure. These measurement instruments were required to meet the following criteria:

- Procedures used to assess validity and reliability,
- Use in empirical studies/citation count, and
- Practicality/feasibility.

The complete list of the final measures can be found in Table 1.

Table 1

Final List of Cognitive and Dispositional Measures Used in the Survey

List of tasks	Construct represented	Construct definition
Fletcher cognitive flexibility	Cognitive complexity	A tendency to understand constructs using a greater number of individual components (i.e., un-simplified). In other words, greater cognitive complexity is associated with more highly differentiated depictions of constructs in memory (Bieri, 1955).
Paper folding	Visualization	Ability to imagine how something will look when it is moved around or when its parts are moved or rearranged (Fleishman, Buffardi, Allen, & Gaskins, 1990).
Zimbardo Time Perspective Inventory	Time perspective	Psychological perception that orients people towards either the past, present, or future; a big-picture perspective: the tendency to understand present events in terms of how they might eventually impact future ones (Zimbardo & Boyd, 1999).
Gestalt completion	Speed of closure	Degree to which different pieces of information can be combined and organized into one meaningful pattern quickly (Fleishman, Quaintance, & Broedling, 1984).
Adaptive expertise	Adaptive thinking	The ability to critique a situation in order to identify potential problems… and generate a set of alternative actions (Joung, Hesketh, & Neal, 2006).
Martin and Anderson Cognitive Flexibility	Cognitive flexibility	Acknowledgement of possible adjustments based on situational factors (Martin & Rubin, 1995).
Analysis-Holism Scale	Holistic thinking tendency	Tendency to think that every element of a task may be interconnected and cannot be understood in isolation from the whole (Choi, Dalal, Kim-Prieto, & Park, 2003).
Conscious presence	Situation awareness	The perception of the environmental elements within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future (Endslev, 1995).
Measures of ability to form spatial mental imagery	Pattern recognition	The process [in] which a person distinguishes a pattern he [or she] percepts with others and identifies what it means (Pi, Liao, Liu, & Lu, 2008).
Word series	Inductive reasoning (fluid intelligence)	Fluid intelligence: Ability to think logically and identify underlying patterns and relationships in novel problems (Ren et al., 2013). Inductive reasoning: Ability to combine separate pieces of information...to form general rules or conclusions (Fleishman, Quaintance, & Broedling, 1984)
Cognitive reflection questions	Analytic thinking	Analytic thinking involves understanding a system by thinking about its parts; The tendency to understand the behavior of an object in terms of cause-and-effect, applying logic and reasoning to predict and explain outcomes (Nisbett, Peng, Choi, & Norenzayan, 2001).
Letter sets	Inductive reasoning (fluid intelligence)	Fluid intelligence: Ability to think logically and identify underlying patterns and relationships in novel problems (Ren et al., 2013). Inductive reasoning: Ability to combine separate pieces of information...to form general rules or conclusions (Fleishman, Quaintance, & Broedling, 1984)
Operation Span Task (OSPAN)	Working memory	Ability to block out irrelevant information and control attention (Engle, 2002); A system that provides temporary storage and manipulation of information needed for complex tasks such as language comprehension, learning, and reasoning (Baddeley, 1986).

MATB-ST Adaptation and Development

Description of Air Force Multi-Attribute Task Battery (AF-MATB)

One major task for us was to select and refine a microworld environment that we could adapt to develop a measure of ST. An effective microworld environment would be both skill-based and domain-general. A skill-based assessment is vital, as it is exponentially more difficult for an individual to intentionally skew results towards the desired outcome. A domain-free measure of ST was sought because it does not limit the assessment to those individuals who are already training in a particular occupation or organization. It was also necessary to simulate our four process-level variables in a microworld environment. Our research, driven by these principles, allowed us to make a shift from a domain-dependent, knowledge-based self-report assessment to a domain-free, skill-based, behavioral assessment of ST.

To do this, we chose the Air Force Multi-Attribute Task Battery (AF-MATB, Miller, Schmidt, Estepp, Bowers, & Davis, 2014; Miller, 2010), a modifiable computer-based simulation, because of its domain-free nature and the ease by which the tasks can be made to be interconnected and yield data on component subtasks. The AF-MATB provides the user with six windows displaying a set of subtasks and usable resources that the user must operate and maintain simultaneously while his or her performance is tracked and recorded (Miller et al., 2014).

The AF-MATB was originally developed as the Multi-Attribute Task Battery by NASA to examine cognitive load and multitasking ability (Comstock & Arnegard, 1992), though it has been used in general decision-making and systems research. For example, since its inception, the MATB has been used to examine physiological driven adaptive task allocation (Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2003), the influence of cognitive abilities on decision-making, the performance implications of goal-setting and mental effort regulation (Venables & Fairclough, 2009), and the effect of workload transitions on neurophysiological states (Bowers, Christensen, & Eggemeier, 2014).

As part of the current research, we mapped the features of the AF-MATB onto the components of ST that we have identified and then modified the simulation to create the MATB-ST (Appendix B shows which systems-thinking subconstructs each MATB-ST subtask measures). Thus, it is not the specific features of the MATB that are of interest, but rather that the MATB requires the processes that are part of our ST conceptual model. A schematic of the MATB-ST can be viewed in Figure 3.

The original simulation consists of four subtasks. The scheduling and pump status windows (no longer a part of MATB-ST) could be set to inform the operator of the current status of the system, as well as the future behavior of the system (note that all of these subtasks can be turned off or automated). The four active subtasks require management of different aspects of the system as a whole. The system monitoring subtask requires the operator to monitor gauges for the occurrence of a system failure. The communications subtask requires the operator to monitor for crucial communications (once received, the operator has to take action to adjust frequencies). The goal of the resource management subtask is to balance and maintain the two consumption tanks (Tank A and Tank B) at a specific volume, using the eight pumps and four supply tanks. The tracking subtask requires the use of a joystick to keep the circular crosshair centered within

the larger crosshair. Most of the time, it is easy to do this, except when there are occasional periods of turbulence. During these periods of turbulence, the location of the circular crosshair is substantially more difficult to control, as it tends to move very quickly and typically outside of the confines of the larger crosshair. Participants with higher levels of systems thinking will be able to pick up on particular task interdependencies and use that information to achieve a higher level of performance on some of the subtasks.

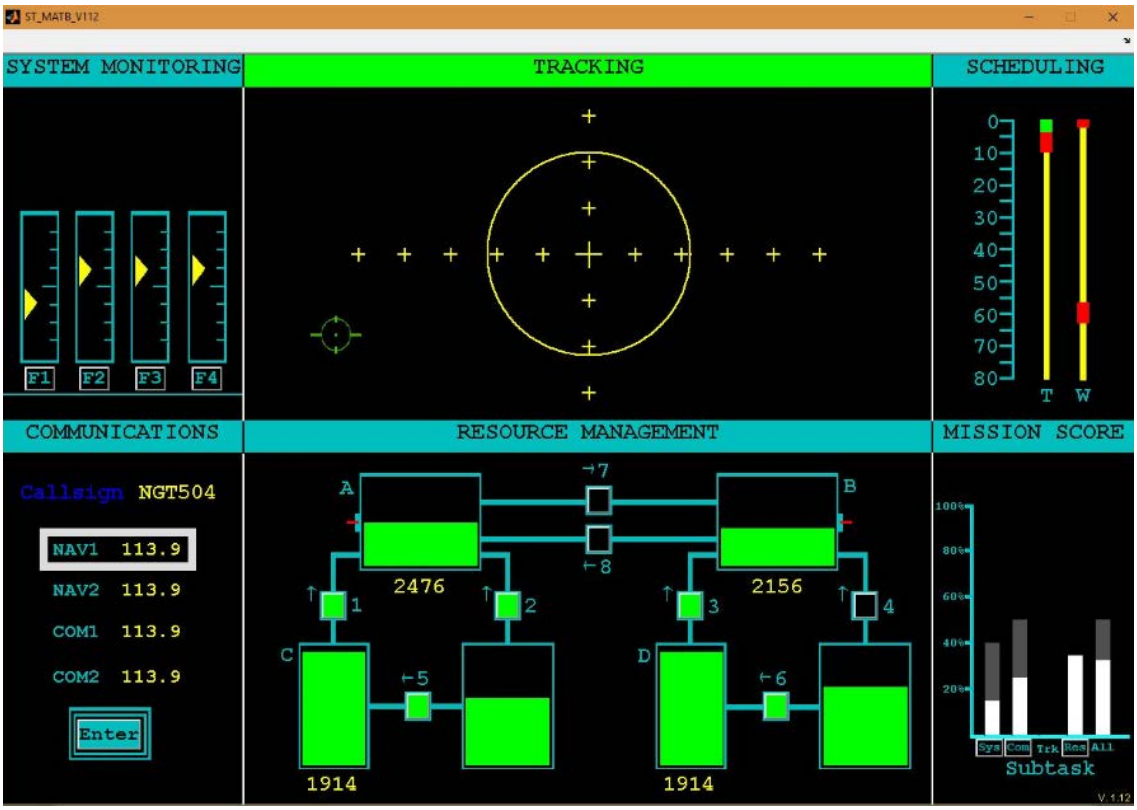

Figure 3. Screenshot of the MATB-ST. In this screenshot, automation for the tracking subtask has been engaged.

Modification of the AF-MATB to the MATB-ST

Through an iterative testing and development process, we modified the original AF-MATB framework to create the MATB-ST (MATB-ST, Figure 3) as our skill-based, domain-general assessment of ST. The following is a description of the end state of the ST constructs after modifications were made to the original framework:

1. The system monitoring subtask (upper left) requires the operator to monitor gauges for a system failure and repair the malfunction. Performance is assessed via the percentage of malfunction gauges correctly identified.

2. The communications subtask (lower left) requires the operator to monitor communications, select the channel, and press [Enter]. Performance is assessed via percent correct.

3. The goal of the resource management subtask (lower middle) is to balance and maintain the two consumption tanks (Tank A and Tank B) between 2,000 and 3,000 liters in volume, using the pumps and supply tanks. Performance is assessed via the proportion of time both tanks are appropriately balanced.

4. The tracking subtask (upper middle) requires the use of a joystick to keep the green reticle within the yellow circle, with performance calculated via the proportion of time participants were successful. The scheduling window (upper right) is linked to the tracking subtask as it provides information regarding upcoming turbulence in order to plan for and address these periods of difficulty by engaging automation. Participants are also able to engage the automation in the absence of turbulence events.

5. The mission score (lower right) provides real-time performance feedback on each subtask and on the MATB-ST as a whole.

We would like to note that the approach we took was one that involved iterative testing and development of the MATB-ST.

MATB-ST Performance Metrics of Systems Thinking

As previously stated, the modifications made to the AF-MATB platform to create the MATB-ST were made specifically to assess the four components of ST defined by our literature review. Systems thinking operationalizations were decided by consensus among a team of human factors and I/O psychologists. The following gives a broad overview of the operationalizations for each component.

1. Holistic thinking is measured by the participant's overall ability to maximize his or her score for each epoch by responding to malfunctions (system monitoring), communications, and to maintain subtask elements within normal ranges (tracking and resource management).
2. Closed-loop thinking is indicated by the participant's performance on the resource management task. This is a measure of closed-loop thinking because it contains all of the elements that closed-loop thinking requires: an understanding of causal relationships, an understanding of the interrelationships of components of a system, and a responsiveness to feedback.
3. Forecasting is measured by two aspects of the scheduling of the automation: (a) the ability to correctly manage the turbulent events, and (b) engaging automation in the absence of a turbulent event.
4. Adaptive/flexible thinking is measured by the speed at which the participant recognizes changes in task priority from one epoch to the next.

MATB-ST Experimentation

This section describes two studies that were conducted. The first study describes data collected for the purposes of validation and development of the MATB-ST. We conducted a second study on Amazon's Mechanical Turk (MTurk), an online data collection platform that gives researchers access to an on-demand, scalable workforce, who complete online surveys in exchange for payment. We conducted this study to attempt to further explore our conceptual model and to ensure that our MATB-ST concept operationalizations measure the ST constructs they were intended to measure. In the section that follows, we will first describe the methodology and results of the MATB-ST study, followed by the description and the results of the MTurk study.

MATB-ST Content Validation Procedure

Before participants arrived, they were asked to fill out informed consent online and all of the survey items in Table 1, which took up to 4 hours to complete. If they had not been completed, participants were given informed consent in the lab and were asked to subsequently complete the survey items at home. Once a participant had been assigned a Participant ID, he or she began the Operation SPAN (OSPAN) working memory task. The participants read through the OSPAN instructions in the lab and were given the opportunity to ask questions.

Upon completion of the OSPAN, participants were then trained on the MATB-ST. Training was primarily conducted in PowerPoint and practice missions of the individual MATB-ST components. The PowerPoint slides were equipped with videos and instructions about how to complete each subtask. After each PowerPoint presentation, participants completed a series of training missions. Mission 1 was the communication and resource management tasks only. Mission 2 was the tracking task and the system monitoring task. During this mission, participants learned how to engage the automation of the tracking task at the appropriate times. If participants did not engage the automation, the tracking task became extraordinarily difficult (indicated by when the red box on the upper right reaches the top; Figure 3). Participants also needed to consider the warm-up and cool-down times for the automation. During Mission 3, participants were trained on all four tasks. The performance assessment began during this period. Participants' scores were penalized for randomly hitting buttons. In Mission 4, participants also trained on all four tasks and then were introduced to the deprioritization of subtasks. During each epoch (the simulation is divided into epochs, which each last 2 minutes), one of the tasks was deprioritized, and the participant was required to identify that task as quickly and accurately as possible. Mission 5 was the final testing mission; it consisted of six 2-minute epochs, for a total of 12 minutes. Performance during Mission 5 determined scores for each of the ST constructs as well as an overall ST score. A subtask was deprioritized during each epoch. The entire testing session (not including time to complete the survey), lasted 75 to 90 minutes. The variability in time is attributable to differences in the time taken to adequately train participants.

Survey demographics. The final sample size was 58 participants. The majority of participants were in their early- to mid-20s (19–59, M = 24.2, SD = 5.8) and evenly divided between male and female (31 of 58 male; 52%). Most were right-handed (55 of 58; 92%), and currently enrolled as students (81%). Consistent with their status as students, the majority of participants reported receiving at least some college education (78%), but a few either had only a

high-school education (19%) or did not specify their level of education (2%). Most participants were White (36%) or Asian (43%), with United States citizenship (74%). The majority also reported English as their first language (63%). Table 2 through Table 5 present detailed demographic data on our sample.

Table 2

MATB-ST Study Education

Education	Number
High school	9
Some college	25
College graduate	11
Post graduate	10
Other	3
No answer	1

Table 3

MATB-ST Study Vocation

Vocation	Number
Craft work	1
Office or clerical	2
Professional	6
Service worker	2
Student	47

Table 4

MATB-ST Study Racial Identification Group

Group	Number
Asian	25
Black	4
Hispanic	3
Other	5
White	21

Table 5

MATB-ST Study Nationality

Country	Number
India	8
Other	7
US	43

MATB-ST results. Table 6 includes the descriptive statistics for the survey data. The majority of the scales collected on Qualtrics had acceptable internal consistency.

Table 6

Descriptive Statistics for Survey Data

Scale	Mean	*SD*	Min	Max	α	Abs. Min	Abs. Max
Analysis holism scale	4.87	.52	3.67	6.08	.77	1	7
Adaptive expertise	4.65	.54	3.38	6.93	.83	1	7
Analytic thinking	1.25	1.15	.00	3.00	.66	0	3
Cognitive complexity	4.33	.52	2.96	5.29	.91	1	6
Cognitive flexibility	2.39	.72	1.08	4.75	.85	1	6
Conscious presence of self-control	2.55	.56	1.30	4.00	.83	1	4
Gaming experience	12.90	22.45	.00	102.00	.44	0	168
Gestalt completion	12.31	2.38	6.00	17.00	.68	0	20
Paper folding	5.88	2.76	.00	10.00	.82	0	10
Spatial mental imagery	16.76	16.65	-4.00	46.00	.87	-46	46
Time perspective (future)	3.36	.43	2.25	4.42	.72	1	5
Word series	17.47	5.91	.00	27.00	.91	0	30
Inductive reasoning	-1.31	.82	-2.50	.50	.34	-3.75	15

Table 7 includes the correlations between the raw performance variables.

Table 7

Intercorrelation Matrix for Raw Performance Values

		2	3	4	5	6	7
1.	System monitoring performance	.12	**.28**	**.30**	.20	.20	-.11
2.	Communications performance		.20	-.12	.13	.14	**-.29**
3.	Resource performance			.01	.20	.22	-.04
4.	Tracking control performance				**.45**	**.43**	.02
5.	Tracking automation event performance					**.96**	-.09
6.	Tracking automation time performance						-.12

Note. Bold font indicates $p < .05$.

Here is a description of each performance metric:

- System monitoring performance: Percentage of malfunction gauges correctly identified by the participant.

- Communications performance: Percentage of critical communications correctly identified. Participant must manually select the channel and press [Enter].

- Resource management performance: Percentage of time both of the key consumption tanks (Tank A and Tank B) are appropriately balanced between 2,000 and 3,000 liters in volume using the pumps and supply tanks.

- Tracking control performance: Percentage of time the participant accurately kept the green reticle within the yellow circle, with performance calculated via the proportion of time participants were successful. Note that this is an overall measure of tracking performance and incorporates all metrics associated with tracking and automation.

- Tracking automation event performance: Percentage of time participants successfully engaged automation to overlap a turbulence event. The scheduling window indicates to the participant when a turbulence event occurs; this metric reflects the ability of the participant to automate the tracking subtask to maintain performance on that task during turbulence.

- Tracking automation time performance: Percentage of time the participants engaged automation in general, including during and in the absence of a turbulence event. More specifically, participants could engage automation between turbulence events, effectively maintaining performance on the tracking subtask without having to control

it manually. This would allow them to allocate attention to the other subtasks, potentially achieving higher performance on these tasks.

Note that tracking automation time performance and time and event performance are essentially the same things as they're highly correlated with one another. The significant correlations between tracking control and tracking automation event performance suggest two possibilities. One possibility is that participants' engagement of automation significantly impacted their control scores (keeping the task in the center of the screen). At first glance, this seems relatively obvious, but participants only engaged 58% of all turbulent events and only 12% of possible extra engagements. When you consider that average performance on the control portion of the tracking subtask was 87% (626/720 seconds), and that total engagement time based on ballpark numbers is 63 seconds, there is no way automation engagements alone can account for the relatively high performance. Another interpretation, given the sum of all of the data, is that participants who paid more attention to the tracking subtask also ended up paying more attention to the automation component of that subtask. The significant negative correlation between communications and extra automation time is unusual, but it could be a function of participants reaching a workload ceiling in which they had to choose between the communications subtask and the engagement of extra automation. In other words, it could be a result of task shedding.

Table 8 presents the descriptive statistics for each subtask (raw performance). Raw performance scores for each subtask were not made available to participants at any time. Note that the raw scores in Table 8 represent proportion data (i.e., theoretical range of 0 to 1) and were subsequently used for the analysis of performance data. To provide participants with performance feedback, the raw performance scores were aggregated into task scores, which were displayed to the participants while they performed the simulation (see Table 9). The displayed scores were not used in the analysis of performance data. The displayed score was important for measuring holistic thinking (see above sections) because the participant needed performance feedback to determine how each subtask contributed to overall performance on the task. As a result, the raw performance scores were simplified for the display and were reset after each epoch, which is the reason the displayed scores differ from the raw performance scores. The purpose of resetting the displayed score was that a different task was prioritized at the start of each epoch. As stated earlier, when a different task was prioritized at the start of each epoch, participants should adapt their strategy to better allocate their attention to the tasks that contributed to a higher score.

Table 8

Descriptive Statistics for the Raw Performance on the Task

Performance	Mean	*SD*	Min	Max
System monitoring performance	0.81	0.14	0.27	1.00
Communications performance	0.84	0.32	0.50	1.00
Resource management performance	0.79	0.24	0.02	1.00
Tracking control performance	0.87	0.19	0.15	0.99
% Turbulence events addressed	0.59	0.33	0.00	1.00
% Turbulence time addressed	0.48	0.28	0.00	0.93
% Extra automations	0.13	0.23	0.00	0.78

Table 9

Descriptive Statistics for the Displayed Scores

Score value	Mean	*SD*	Min	Max
System score	67.80	12.19	23.33	83.33
Communications score	70.06	20.03	-6.25	83.33
Resource score	63.55	19.49	2.08	81.26
Total score	201.40	35.07	91.63	245.38

Next, we present the descriptive statistics for each of the four ST components as assessed by the MATB. We have one metric each for holistic thinking (total performance score), adaptive/flexible thinking (correctly answering deprioritized questions) and closed-loop thinking (resource management score), and a measure of forecasting (frequency and accuracy of automation engagements). These descriptive statistics are shown in Table 10.

Table 10

Component Descriptive Statistics

Scale	M	*SD*	Min	Max
Holistic thinking	0.81	0.06	0.12	1.00
Adaptive/flexible thinking	0.53	0.23	0.00	1.00
Closed-loop thinking	0.78	0.11	0.00	1.00
Forecasting	0.54	0.38	0.00	1.00

Relation of MATB-ST and survey data. Table 11 shows simple (linear) mixed-effects results. The betas contained inside can be interpreted like zero-order correlation coefficients but have the added benefit of accounting for nesting effects. The cells for the individual differences variables (and the facets of ST) are color-coded according to their consistency with our priori predictions (see Appendix B). Italicized cells refer to relationships we expected to be positive, and non-italicized cells refer to those we had no predictions for or expected to be non-significant.

Table 11

The Relation of 14 Scales/Ability Metrics With our Four Systems Thinking (ST) Components

	Holistic		Adaptive		Closed−loop		Forecasting	
	β	*p*	β	*p*	β	*p*	β	*p*
Epoch	.01	.736	.00	.97	-.02	.48	.03	.44
Holistic	—	—	*.20*	*.00*	*.54*	*.00*	*.05*	*.32*
Adaptive	*.27*	*.00*	—	—	*.12*	*.01*	*.15*	*.00*
Closed loop	*.59*	*.00*	*.00*	*.97*	—	—	*.01*	*.86*
Forecasting	*.06*	*.29*	*.20*	*.00*	*.01*	*.91*	—	—
Analysis holism scale	-*.07*	*.51*	-.01	.93	-.03	.78	-*.08*	*.48*
Adaptive expertise	.06	.53	-*.08*	*.39*	-.04	.69	.03	.78
Analytic thinking	*.23*	*.03*	*.16*	*.06*	*.27*	*.01*	*.29*	*.01*
Cognitive complexity	.05	.62	*.10*	*.28*	*.08*	*.46*	-*.04*	*.67*
Cognitive flexibility	-.21	.05	-*.13*	*.15*	-*.22*	*.04*	.07	.50
Conscious presence	*.11*	*.27*	-*.06*	*.49*	-*.10*	*.35*	*.07*	*.54*
Gaming	.09	.41	.06	.50	.09	.40	-.10	.38
Gestalt completion	-*.01*	*.92*	.03	.72	.06	.56	-*.26*	*.01*
Paper folding	.10	.35	.10	.27	*.13*	*.24*	.15	.15
Spatial mental imagery	*.22*	*.03*	*.06*	*.52*	*.23*	*.03*	.11	.30
Time perspective	.09	.37	*.03*	*.73*	-.02	.86	-*.24*	*.02*
Word series	.20	.05	.10	.25	.28	.01	-.06	.59
Inductive reasoning	-*.01*	*.89*	-*.03*	*.77*	-*.01*	*.91*	-*.12*	*.28*
OperationACC	.12	.26	.11	.22	.18	.10	.17	.11
OperationRT	-.09	.37	.01	.89	-.02	.87	.00	.99
MathACC	.12	.27	.13	.14	-.06	.57	.12	.26
MathRT	-.19	.06	-.09	.32	-.06	.56	-.16	.13

Note. The data above the dashes represent the intercorrelations among the four components. OperationACC, OperationRT, MathACC, and MathRT are all metrics from the OSPAN. Cells italicized refer to relationships we expected to be positive, and non-italicized cells refer to those we had no predictions for or expected to be non-significant.

General conclusions of the MATB-ST Study. The goal of data collection conducted with the MATB-ST was two-fold. First, we wanted to get a sense of how participants generally performed with the MATB-ST. Importantly, we wanted to ensure that the collection of tasks did not induce an unreasonable amount of workload. A perusal of the demographics of the MATB-ST subtasks reveals that average performance exceeded 75% in most cases, suggesting that all the subtasks can be performed reasonably well. The data also revealed that participants sparingly used the tools related to the engagement of automation. There are two possible explanations for this finding. First, it could be the case that participants felt they had achieved mastery over this task and didn't recognize an added benefit of using the automation. Indeed, there are studies that show that motor learning can occur quite rapidly, and mastery can be achieved in as little as 30 minutes (e.g., Hill & Schneider, 2006). Another possibility is that participants experienced a cognitive tunneling effect because of the labor-intensive nature of the MATB-ST, thereby limiting any additional resources to allocate to additional tasks.

A couple of points are worth noting. The gaming experience is unrelated to all four of our subscales, which combats any claim that the MATB-ST measure is convoluted by computer skills and motor ability. One exception worth noting is that if all four subscales were high to start with, then the range restriction might cover game experience effects. Also, our measures of working memory and inductive reasoning are unrelated to any of our four components. This suggests that we are tapping a construct that cannot be considered general mental ability or intelligence. Likewise, the same range restriction argument may apply as an alternate explanation about why our measures of working memory and inductive reasoning were unrelated to the four ST components.

There were also some surprising and perhaps counter-intuitive findings in the current set of data. For example, the MATB-ST derivation of holistic thinking is unrelated to scores on the analysis holism scale. Additionally, the MATB-ST derivation of adaptive/flexible thinking is unrelated to scores on the cognitive flexibility scale. A possible explanation for these findings could be that the cognitive and dispositional-type individual differences measures don't map on well to our ability-based assessment of ST. For example, common measures of cognitive flexibility are ability-based metrics that require task and goal switching (e.g., Youmans, Figueroa, & Kramarova, 2011). Perhaps a measure that taps cognitive flexibility and holistic thinking as an ability, and not self-report or disposition, would produce better correlational mappings between those predictors and the MATB-ST.

Overall, the general conclusion that can be drawn is that the MATB-ST is showing promising correlations with our variables. The logical next step is to examine how the MATB-ST predicts a criterion such as job performance. Next steps and future directions are discussed under General Discussion, beginning on page 29.

ST Criterion Validation Research Using an MTurk Sample

We conducted an additional study. The purpose of the second study was to examine convergent validity, as well as criterion-related validity, by examining extant measures of ST and how adept current measures of ST are at predicting job performance. The research on ST to date has been primarily conceptual, and the majority has focused on how leaders' usage of ST can improve organizational performance (Davis et al., 2015; Dzombak, Mehta, Mehta, & Bilén, 2014). However, empirical work that connects ST as an individual difference to individual-level outcomes (e.g., job performance) is important to empirically demonstrate the benefit of ST for selection across a variety of jobs.

In this follow-up study, we examined the effects of two domain-general conceptualizations: one in which ST is viewed more as an ability/skill (i.e., the ability/skill to "represent and assess behavior that arises from the interaction of a system's agents over time"; Sweeney & Sterman, 2000, p. 250), and another in which ST is viewed more as a dispositional tendency/preference (i.e., "an implicit tendency to recognize various phenomena as a set of interconnected components that interact with one another to make up a dynamic whole"; Davis & Stroink, 2015, p. 3). We assessed the impact of these two conceptualizations of ST together with our other cognitive and dispositional individual differences (see Table D-2 in Appendix D). This allowed us to examine whether ST mediates the effects of these variables on, and explains incremental variance beyond, these variables in job performance.

We hypothesized that ST will relate positively to job performance. Individuals higher in ST are more likely to identify connections between work task components and to consider them from a dynamic and holistic view. These individuals should understand and solve work-related problems more successfully and, therefore, achieve higher job performance. It has previously been found that domain-specific ST benefited performance within corresponding domains (Davidz & Nightingale, 2008; Frank, 2006). The present study seeks to evaluate this relationship with ST conceptualized as an individual difference whose effects on job performance generalize across domains. It was also hypothesized that ST will not only explain incremental variance beyond the cognitive and dispositional measures but also mediate the effects of these variables on job performance. Finally, we hypothesized that job complexity will moderate the ST-performance relationship, such that this relationship will be stronger for more complex jobs. The full model can be viewed in Figure 4.

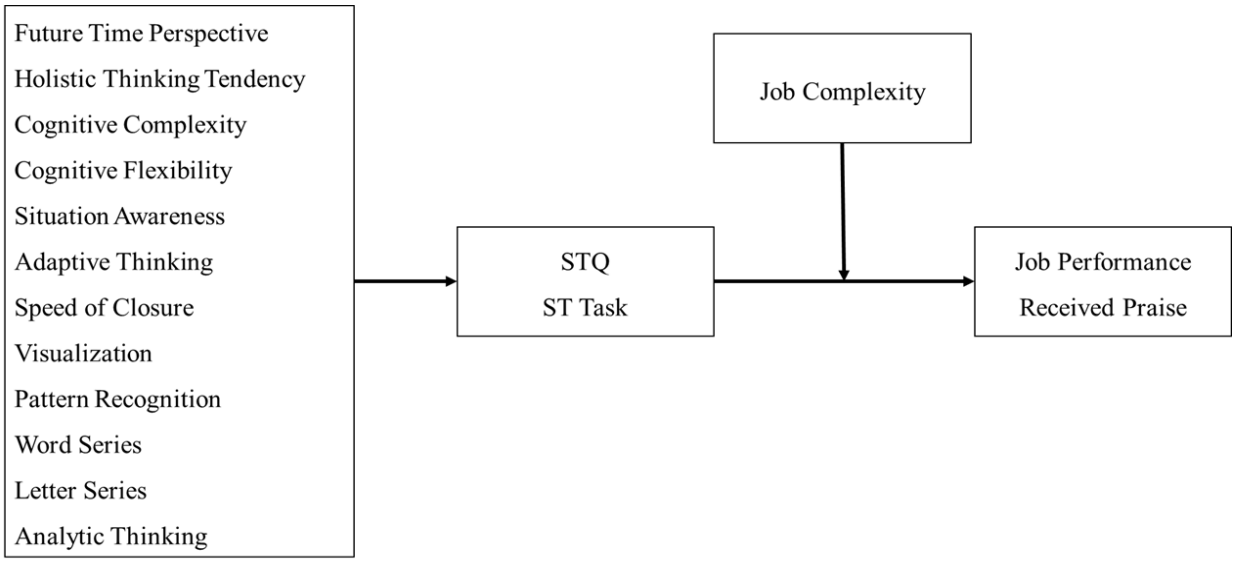

Figure 4. Full model for the criterion validation research.[1]

ST criterion validation research procedure. We recruited 406 participants through Amazon.com's Mechanical Turk, of which 49.6% were female, and the ages of all participants range from 18 to 45. All participants were employed in the U.S. and had held only one job (position) at one organization in the past year (these restrictions were felt to be necessary to give employees the potential to develop an understanding of the systems associated with their job). We used two extant measures for ST: the 15-item ST Questionnaire (STQ; Davis & Stroink, 2015), in which ST is conceptualized as a dispositional tendency/preference, and the ST Task that included four graph-reading questions (Sterman, 2002), in which ST is conceptualized as an ability or skill. The ST Task measured basic ST concepts such as stocks and flows. Job performance was measured using a single item and, to ameliorate concerns about the validity of self-rated measures of performance, using several biodata-like (and hence open to less subjective interpretation) items about how often employees were praised for their performance in the last year (see Table 12).

Note that numeracy (Peters, Dieckmann, Dixon, Hibbard, & Mertz, 2007) was measured as a control variable. Job complexity was computed from three sets of questions: O*NET Job Zones questions (see https://www.onetcenter.org/), job requirements of the systems skills, and managerial status (see Table 12).

[1] STQ = 15-item ST Questionnaire (Davis & Stroink, 2015) in which ST is conceptualized more as a dispositional tendency/preference. ST Task = 4 graph-reading questions (Sternman, 2002) measuring basic ST concepts (e.g., stocks and flows, time delays) in which ST is conceptualized more as an ability or skill.

Table 12

Items Used for Measuring Moderator and Outcome Variables

Variable	**Items**
O*NET Job Zones (multiple choices)	What is the highest level of education required to do your job? Does someone need to attend vocational school, have related on-the-job experience, or receive some kind of specialized training to do your job? How much experience does someone need to do your job? How much formal training after being hired is required to do your job?
Job Requirements of Systems Skills [a] (5-point Likert scale)	*Indicate how often your job requires you to do the following:* - Consider the relative costs and benefits of potential actions to choose the most appropriate one - Determine how a system should work and how changes in conditions, operations, and the environment will affect outcomes - Identify measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system
Managerial Status	How many employees do you manage? How many managers do you manage? How many units/departments do you manage?
Job Performance (7-point Likert scale)	On average, how well do you perform the core tasks associated with your job? Please compare your performance to that of the average person who holds a similar job in your organization.
Received Praise (5-point Likert scale)	*In the past year, how many times have you:* - Been praised by your supervisor because of your performance on one or more projects? - Received a very positive formal performance review? - Finished projects more quickly than your peers? - Received complaints about your performance on one or more projects? (R) - Been taken off one or more projects because of inadequate performance? (R)

Note. (R) indicates reverse-coded items. [a]We used systems skills listed on O*NET (see https://www.onetcenter.org/).

ST criterion validation research results. All results from the MTurk criterion validation research are included in Appendix D. What follows is a brief description and interpretation of the results.

Simple regression analyses revealed that after controlling for numeracy, the 15-item dispositional preference/tendency measure of ST (also known as the ST Questionnaire or STQ) was a significant predictor of job performance ($\beta = .17$, $p < .01$) and received praise ($\beta = .25$, $p < .01$). In contrast, the graph-reading ST Task (the ability/skill measure of ST) was not a significant predictor. Therefore, hierarchical regression and moderation analyses were conducted only for the STQ.

Hierarchical regression analyses revealed that STQ demonstrated incremental validity in relation to job performance and received praise over and above most of the antecedents. Mediation analyses showed that STQ was a partial mediator for all the significant antecedents of received praise, and for half of the significant antecedents of job performance, suggesting that ST was an important mediator between individual differences and performance. In contrast, the ST Task was not a significant mediator of any relationship.

Moderation analyses revealed that job complexity was not a significant moderator of either the STQ-Job Performance or the STQ-Received Praise relationship. Due to the large sample size, the absence of significant moderation results cannot be attributed solely to low statistical power.

General conclusions: ST criterion validation research. The results of this follow-up study provided empirical evidence of the importance of ST for job performance at the individual employee level. In other words, ST can predict performance over and above more distal individual differences, but also, as a mediator, can partially explain the effects of these antecedents. In addition, the effects of ST did not differ between complex jobs and simple jobs, suggesting that ST is equally important for a variety of job types, even those that are low in complexity and that do not, on their face, appear to require "systems skills." These results provide preliminary support for the idea that a measure of ST could be used to select employees for a variety of jobs. Moreover, if ST is malleable (i.e., a skill rather than an ability), training ST (e.g., via mental models; Senge, 1990) could be useful for a variety of jobs.

Considering the disadvantages (such as fakability; King & Bruner, 2000) of self-report disposition measures (e.g., STQ), a skill/ability (or competency) measure of ST is at least equally needed. However, the skill/ability measure we used (i.e., the ST Task) did not predict performance-related outcomes. Such a skill-based measure should assess multiple components of ST (e.g., holistic thinking vs. forecasting) and should probably involve a larger "system" with more opportunities to demonstrate ST (e.g., multiple trials, multiple ST-relevant responses per trial). This also speaks to the need to develop alternative skill-based measures of ST, which was the goal of the MATB-ST study.

General Discussion

Although ST has been a focus of systematic study across multiple disciplines over many years (Frank, 2012; Senge, 1990; Smuts, 1926; von Bertalanffy, 1968), a clear conceptual framework and operationalization have been lacking. Accordingly, we sought to (a) refine the conceptual definition and framework of the ST construct; (b) develop a skill-based measure of system-thinking; and (c) provide some initial validation of both the conceptual framework and measure of this construct. The current research used a review of the literature to derive four components of ST: holistic thinking, adaptive/flexible thinking, closed-loop thinking, and forecasting. The first project in this effort provided a measure of these components, the MATB-ST. The second project, involving the ST criterion validation research, provided support for the criterion-related validity of ST relative to job performance, as well as some initial confirmation of some proposed antecedents. Taken together, these projects provide a conceptual foundation for future research on ST, as well as a possible skill-based measure to assess this construct. The present and ensuing research also offers a number of significant potential applications for the U.S. Army. In this section, we discuss future research possibilities, as well as implications of our findings in this project for the Army. Before doing so, though, we summarize some limitations in this work.

Limitations

Our findings from both studies may be limited in some ways. The MATB-ST study was completed primarily with college students. Although the age range for this sample parallels that of Army recruits, some may argue that this sample is more skilled in computer-based games than older samples, skewing potential results. Our measure of gaming experience was not correlated with any of the component scores, suggesting that such experience was not a bias. However, future research should test the MATB-ST with a more variable sample, as a more homogeneous sample with most individuals having gaming experience is another possible reason why we did not see any relationship with the component scores. Also, given its intended application in the Army, it should also be tested with a range of military samples.

The results of the ST criterion validation research should be treated with caution, as it is susceptible to common method variance because of its self-report methodology. Such a bias would raise the probability of observing direct and mediated relationships among measured variables. We would note that some researchers have argued that common method variance may be reduced when participants know their responses are confidential (Aquino, Galperin, & Bennett, 2004), as was the case in this study. Also, while our measures of performance in this study were both self-report, one set of measures (“received praise”) was a biodata assessment. Kilcullen (1993) reported that such biodata measures were as accurate, or more so, than objective records. Nonetheless, future studies that seek to replicate these results should use a longitudinal design to assess predictors, mediators, and outcomes of ST. Thus, the current ST criterion validation research project should be viewed as a preliminary demonstration that ST can predict job performance.

Future Research

This project has yielded promising findings that provide a foundation for several future research directions. We will enumerate in this section some of the directions we feel are most critical. These directions concern both the conceptualization of ST and the continued validation of the MATB-ST.

Our framing of the systems thinking construct argues for four components. The results of the MATB-ST study indicated that the proposed measure provides distinct assessments of each component. As is the case with the structure of the ST model, further research will need to provide additional validation for these components as necessary elements of ST. Doing so will require a demonstration of (a) the convergent validation of the components themselves, as well as (b) their relatively equal relationships with other assessments and outcomes of ST. The ST criterion validation research study provided some initial evidence in that several measures that were related to the four components as antecedents (i.e., future time perspective, holistic thinking, cognitive flexibility, adaptive thinking, and situation awareness) were also significantly related to measures of job performance. Furthermore, four of these five measures were related to the dispositional measure of ST. Future research will need to test whether the four subscales of ST that we identified are necessary and sufficient for an individual to possess ST capacity. In other words, subsequent research might test whether the absence of any one component means the lack of the overall capacity. Such tests would involve evaluating the measure against ST criteria and noting whether the multiplicative combination of the components provides greater predictability than their additive combination.

We have also argued that the components of ST represent skills that are relatively malleable and can be developed through focused learning activities. However, other researchers have argued that constructs related to ST, such as conceptual capacity (Jaques, 1986, 1989), are relatively immutable, with change only possible within certain limited capacity bands. Future research will need to explore further the question of the malleability of ST. Such studies may entail the testing of particular instructional strategies designed to enhance each component and examine evidence of possible growth. A key research question for such research would be whether some but not all of the components are malleable. Evidence of differential malleability would have significant implications for training and development programs targeting ST. The malleability of the ST components would also indicate whether the ST assessment would best be used for selection (if the components are non-malleable) or within a training context. Research on how much of ST reflects malleable skills or immutable abilities is necessary before implementing efforts to develop ST.

The ST criterion validation research study included measures of job performance as criterion variables for ST. The results were promising, indicating evidence of its criterion-related validity. However, future research is necessary to (a) specify more clearly what kinds of performance are most likely to be predicted by ST capacities (e.g., might forecasting skill lead to proactive behavior?), and (b) determine what types of work tasks are most likely to yield higher predictive validities of ST with performance. Our assessment of job performance included a self-report measure of general job performance and biodata-type measures of received praise for several performance-related outcomes. These are distal outcomes of ST. Future research should

provide additional tests of the effects of ST capacity with such distal outcomes. However, such research should also examine more proximal outcomes that may be related to complex problem solving. We would argue that ST should facilitate several processes associated with complex problem solving, including situation scanning and awareness, sense-making, solution generation, solution evaluation, and solution implementation planning (Byrne, Shipman, & Mumford, 2010). Studies to test this assertion will need to carefully parse a skill-based measure such as the MATB-ST from other indicators of complex problem solving. We would expect that ST capacity demonstrated on the MATB-ST should be associated with effective performance in other complex problem domains, as evidenced by solution speed and solution quality. Such research would also test our assumption that the MATB-ST represents a domain-general assessment of ST. For similar reasons, we believe that ST capacity should also predict other outcomes related to complex problem solving such as creativity, innovation, and adaptive performance (Byrne et al., 2010; Zaccaro, Weis, Chen, & Matthews, 2014).

Work tasks vary in terms of their requisite complexity. Tasks that are ill-defined, with multiple components and dynamic elements, are more complex than well-defined and static tasks. The components of ST should be more necessary for performance on more complex tasks (Jacobs & Jaques, 1987; Zaccaro, 2001). We would argue, therefore, that job complexity should moderate the relationship between ST and job performance. Note that the MTurk study did not provide support for this assertion. However, our measures of job complexity in that study were self-report. This raises the question of whether job incumbents could reasonably construe the complexity of their jobs and whether such scores were susceptible to perceptual biases. Thus, future research will need to use more objective measures of job complexity to assess whether it acts as a moderator of the effects of ST.

The ST criterion validation research study examined a range of cognitive and dispositional individual differences as potential antecedents of ST. However, prior research has suggested that additional personality and motivational attributes may also be important precursors to ST (Mumford, Zaccaro, Harding, Jacobs, & Fleishman, 2000). Such qualities as drive, ambition, and achievement motivation would predispose individuals to engage the more effortful cognitive resources associated with ST. Likewise, dispositional attributes such as tolerance for ambiguity and openness to experience would provide a foundation for proactive engagement in the complex task situations that require ST. Accordingly, future research should expand the nomological network of variables predicting ST beyond those examined in the ST criterion validation research study.

The present research provided promising evidence for MATB-ST as a viable assessment of ST capacity. However, several key questions remain. First, additional evidence is needed to support the construct validity of the MATB-ST. Preliminary evidence from our effort suggests modest convergence with some related constructs. Measurement differences and potential self-report biases likely attenuated these validity coefficients. Accordingly, future research will need to construct a tighter network of measures that can provide a more accurate assessment of the respective convergent and divergent validity of the MATB-ST. Additional work is also needed to further refine and improve the MATB-ST operationalizations of the four components of ST. Such studies should be completed with both non-Army and Army samples, as the former can provide a basis for refinement of the MATB-ST before going to an Army sample.

Another research question pertains to whether the MATB-ST outperforms other measures of ST in predicting targeted criteria. Accordingly, further research will need to examine the criterion-related validity of the MATB-ST. The current effort provided evidence from the ST criterion validation research study that ST is a predictor of job performance. However, the MATB-ST study did not provide a similar assessment. For the MATB-ST to be a viable assessment tool for the Army, future studies will need to define both proximal and distal performance criteria and demonstrate significant criterion-related validities in an Army sample. Such studies will also need to demonstrate that the MATB-ST can provide stronger validities than current measures of ST. This research direction will also allow tests of research questions raised earlier, including whether (a) the four components of ST assessed by the MATB-ST are equally predictive of performance, and (b) whether job complexity is a moderator of such validities.

Also, future research may examine the extent to which the MATB-ST is a measure of more than simply multi-tasking ability. However, this may be challenging, given that the MATB necessarily is measuring multiple ST constructs at the same time.

Finally, after establishing the construct and criterion-related validity of the MATB-ST, future efforts should turn to developing the web-enabled operational version for use in Army mass assessments. This version will need to demonstrate acceptable practicality (e.g., not requiring overly burdensome time commitment; fitting within current Army mass assessment protocols; engaging to participants) with no significant reduction in construct and criterion-related validities over the prototype version.

Implications for the U.S. Army

A validated skill-based measure of ST can provide significant benefits to the U.S. Army. A key question is its utility for selection as well as classification. If subsequent studies demonstrate that success on the MATB-ST is associated with higher performance across a broad spectrum of Army military occupational specialties (MOSs), then this measure may conceivably be used as an entry level selection tool. However, we would still argue that the MATB-ST is likely to be more predictive of performance in more complex jobs. Accordingly, this measure can be valuable as a classification tool for Army MOSs that are rated higher in job complexity.

We would add that job complexity refers to both information and social complexity (Zaccaro, 2001). Thus, for example, MOSs related to cyber security, intelligence, and strategic planning may all exhibit higher levels of informational complexity. Alternatively, MOSs related to foreign affairs, civil affairs, joint operations, and others that entail extensive liaising and interactions with external Army stakeholders will likely have higher levels of social complexity. Both forms of complexity require a greater understanding of systems, albeit of different kinds. A key research objective should be validating the MATB-ST under both kinds of complexity. If such evidence is procured, then the MATB-ST may serve as a valuable classification tool for MOSs possessing both forms of job complexity. It is also important to note that the goal of this project was to develop a domain independent measure of ST rather than a domain specific measure of ST. A domain independent measure of ST would have far more value to the Army

than a domain specific measure of ST, as it could be applicable across a wide variety of MOSs where ST is important.

To the extent that research on the MATB-ST indicates that one or more of the components of ST are malleable, this would have significant implications for Army training and development. If all four components are malleable, then instructional protocols for their development could be constructed and applied in the Army for Soldiers entering certain MOSs (assuming that the MATB-ST is a stronger predictor of performance in some MOSs than in others). If any of the components of the ST are immutable, then assessments of those components may serve as assessments of developmental readiness for training on the more trainable elements of ST.

Finally, a validated MATB-ST measure may be a valuable tool for officer selection, classification, and development. Several researchers have argued that informational and social complexity are key elements of leadership positions and that such complexity increases as leaders ascend organizational ranks (Jacobs & Jaques, 1987; Mumford, Campion, & Morgeson, 2007; Mumford, et al., 2000; Zaccaro, 2001). This argument suggests that measures can be used as potential selection and classification tools for entry noncommissioned and commissioned officer positions. Alternatively, a potentially powerful use of a ST measure could be to provide officers with a tool to self-assess their capacity for ST and have them use the results to compare and contrast officer development plans that target growth in overall ST or in one or more of its components.

Conclusion

Historically, ST has become an increasingly more critical work skill for personnel to have in the U.S. Army. However, the construct has suffered from an imprecise definition and inadequate measurement. The projects described in this report provide a foundation for improving the conceptualization, operationalization, and assessment of this construct. Future research that builds on this foundation can be of substantial benefit to understanding and measuring ST.

References

Adam, T., & de Savigny, D. (2012). Systems thinking for strengthening health systems in LMICs: need for a paradigm shift. *Health Policy and Planning*, *27*(suppl 4), iv1–iv3. https://doi.org/10.1093/heapol/czs084

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, *76*(5), 732–740. https://doi.org/10.1037/0021-9010.76.5.732

Aquino, K., Galperin, B. L., & Bennett, R. J. (2004). Social status and aggressiveness as moderators of the relationship between interactional justice and workplace deviance. *Journal of Applied Social Psychology*, *34*(5), 1001–1029. https://doi.org/10.1111/j.1559-1816.2004.tb02581.x

Baddeley, A. (1986). *Working memory.* Oxford, England: Clarendon Press.

Bahner, J., Hüper, A., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Bieri, J. (1955). Cognitive complexity-simplicity and predictive behavior. *The Journal of Abnormal and Social Psychology*, *51*(2), 263–268. https://dx.doi.org/10.1037/h0043308

Bowers, M. A., Christensen, J. C., & Eggemeier, F. T. (2014). The effects of workload transitions in a multitasking environment. *Proceedings of the Human Factors & Ergonomics Society 58th Annual Meeting*, *58*(1), 220–224. https://dx.doi.org/10.1177/1541931214581046

Bransford, J. D., & Franks, J. J. (1976). Toward a framework for understanding learning. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 10, pp. 93–127). New York, NY: Academic Press. http://dx.doi.org/10.1016/S0079-7421(08)60465-X

Byrne, C. L., Shipman, A. S., & Mumford, M. D. (2010). The effects of forecasting on creative problem-solving: An experimental study. *Creativity Research Journal*, *22*(2), 119–138. https://dx.doi.org/10.1080/10400419.2010.481482

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://dx.doi.org/10.1037/h0046016

Cavaleri, S., & Sterman, J. D. (1997). Towards evaluation of systems thinking interventions: A case study. *System Dynamics Review*, *13*(2), 171–186.

Chapanis, A. (1996). *Human factors in systems engineering*. New York, NY: Wiley.

Checkland, P. (1997). Systems thinking. In W. L. Currie & B. Galliers (Eds.), *Rethinking management and information systems: An interdisciplinary approach* (pp. 45–56). New York, NY: Oxford University Press.

Choi, I., Dalal, R., Kim-Prieto, C., & Park, H. (2003). Culture and judgement of causal relevance. *Journal of Personality and Social Psychology, 84*(1), 46–59. https://dx.doi.org/10.1037/0022-3514.84.1.46

Comstock, J. R., Jr., & Arnegard, R. J. (1992*). The multi-attribute task battery for human operator workload and strategic behavior research* (Technical Memorandum No. 104174). Hampton, VA: NASA Langley Research Center.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://dx.doi.org/10.1037/h0040957

Davidz, H. L., & Nightingale, D. J. (2008). Enabling systems thinking to accelerate the development of senior systems engineers. *Systems Engineering*, *11*(1), 1–14. https://dx.doi.org/10.1002/sys.20081

Davis, A. C., & Stroink, M. L. (2016). The relationship between systems thinking and the new ecological paradigm. *Systems Research and Behavioral Science, 33*(4), 575–586. https://dx.doi.org/10.1002/sres.2371

Davis, A. P., Dent, E. B., & Wharff, D. M. (2015). A conceptual model of systems thinking leadership in community colleges. *Systemic Practice and Action Research, 28*(4), 333–353. https://dx.doi.org/10.1007/s11213-015-9340-9

Dzombak, R., Mehta, C., Mehta, K., & Bilén, S. G. (2014). The relevance of systems thinking in the quest for multifinal social enterprises. *Systemic Practice and Action Research 27*(6), 593–606. https://dx.doi.org/10.1007/s11213-013-9313-9

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32–64. https://dx.doi.org/10.1518/001872095779049543

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65–84. https://dx.doi.org/10.1518/001872095779049499

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Fleishman, E. A., Buffardi, L. C., Allen, J. A., & Gaskins, R. C., III (1990). *Basic considerations in predicting error probabilities in human task performance* (No. NUREG/CR-5438). Washington, DC: Nuclear Regulatory Commission.

Fleishman, E. A., Quaintance, M. K., & Broedling, L. A. (1984). *Taxonomies of human performance: The description of human tasks.* Orlando, FL: Academic Press.

Frank, M. (2006). Knowledge, abilities, cognitive characteristics and behavioral competences of engineers with high capacity for engineering systems thinking (CEST). *Systems Engineering*, *9*(2), 91–103. https://doi.org/10.1002/sys.20048

Frank, M. (2012). Engineering systems thinking: Cognitive competencies of successful systems engineers. *Procedia Computer Science*, *8*(1), 273–278. https://doi.org/10.1016/j.procs.2012.01.057

Hill, N. M., & Schneider, W. (2006). Brain changes in the development of expertise: Neuroanatomical and neurophysiological evidence about skill-based adaptations. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 653–682). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.037

Jacobs, T. O., & Jaques, E. (1987). Leadership in complex systems. In J. Zeidner (Ed.), *Human productivity enhancement, Volume 2: Organizations, personnel, and decision making* (pp. 7–65). New York, NY: Praeger.

Jaques, E. (1986). The development of intellectual capability: A discussion of stratified systems theory. *The Journal of Applied Behavioral Science*, *22*(4), 361–383. https://dx.doi.org/10.1177/002188638602200402

Jaques, E. (1989). *Requisite organization: The CEO's guide to creative structure and leadership*. Arlington, VA: Cason Hall.

Joung, W., Hesketh, B., & Neal, A. (2006). Using "war stories" to train for adaptive performance: Is it better to learn from error or success? *Applied Psychology*, *55*(2), 282–302. https://dx.doi.org/10.1111/j.1464-0597.2006.00244.x

Kilcullen, R. N. (1993). *The development and use of rationally-keyed background data scales to predict leader effectiveness* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Database. (Order No. 9407460).

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing*, *17*(2), 79–103. https://dx.doi.org/10.1002/(SICI)1520-6793(200002)17:2<79: AID-MAR2>3.0.CO;2-0

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://dx.doi.org/10.1111/j.1744-6570.1975.tb01393.x

Martin, M. M., & Rubin, R. B. (1995). A new measure of cognitive flexibility. *Psychological Reports*, *76*(2), 623–626. https://dx.doi.org/10.2466/pr0.1995.76.2.623

Milheim, W. D., & Martin, B. L. (1991). Theoretical bases for the use of learner control: Three different perspectives. *Journal of Computer Based Instruction, 18*(3), 99–105.

Miller, W. D. (2010). *The U.S. Air Force-developed adaptation of the Multi-Attribute Task Battery for the assessment of human operator workload and strategic behavior* (Technical Report 2010-0133). Wright-Patterson Air Force Base, OH: Air Force Research Laboratory. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA537547

Miller, W. D., Schmidt, K. D, Estepp, J. R., Bowers, M., & Davis, I. (2014). *An updated version of the U.S. Air Force Multi-Attribute Task Battery (AF-MATB)* (Special Report 2014-0001). Wright-Patterson Air Force Base, OH: Air Force Research Laboratory. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA611870

Moroney, W. F., Biers, D. W., & Eggemeier, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *International Journal of Aviation Psychology, 5*(1), 87–106. https://dx.doi.org/ 10.1207/s15327108ijap0501_6

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191–205. https://dx.doi.org/10.1177/001316444700700201

Mumford, M. D., Zaccaro, S. J., Harding, F. D., Jacobs, T. O., & Fleishman, E. A. (2000). Leadership skills for a changing world: Solving complex social problems. *The Leadership Quarterly, 11*(1), 11–35. https://dx.doi.org/10.1016/S1048-9843(99)00041-7

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2007). The leadership skills strataplex: Leadership skill requirements across organizational levels. *The Leadership Quarterly, 18*(2), 154–166. https://dx.doi.org/10.1016/j.leaqua.2007.01.005

Natsoulas, T. (1967). What are perceptual reports about? *Psychological Bulletin, 67*(4), 249–272. https://dx.doi.org/ 10.1037/h0024320

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291–310. https://dx.doi.org/10.1037/0033-295X.108.2.291

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://dx.doi.org/10.1037/0033-295X.84.3.231

Olszewski, D. H. (2014). *The use of systems thinking by the industrial engineer as organizational leader* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Database. (Order No. 3623746).

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review*, *64*(2), 169–190. https://dx.doi.org/10.1177/1077558707064002030l

Youguo Pi, Wenzhi Liao, Mingyou Liu and Jianping Lu (2008). Theory of Cognitive Pattern Recognition, Pattern Recognition Techniques, Technology and Applications, Peng-Yeng Yin (Ed.), InTech, DOI: 10.5772/6251. Available from: https://mts.intechopen.com/books/pattern_recognition_techniques_technology_and_applications/theory_of_cognitive_pattern_recognition

Prinzel, L. J., III, Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential P300 component. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(4), 601–614. https://dx.doi.org/10.1518/hfes.45.4.601.27092

Richmond, B. (1994). Systems thinking/system dynamics: Let's just get on with it. *System Dynamics Review*, *10*(2-3), 135–157. https://dx.doi.org/10.1002/sdr.4260100204

Sauer, J., Wastell, D. G., & Hockey, G. R. J. (2000). A conceptual framework for designing micro-worlds for complex work domains: a case study of the Cabin Air Management System. *Computers in Human Behavior*, *16*(1), 45–58. https://dx.doi.org/10.1016/S0747-5632(99)00051-5

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization*. New York, NY: Doubleday.

Smuts, J. C. (1927). *Holism and evolution.* New York, NY: MacMillan & Co. Retrieved from https://archive.org/details/holismandevoluti032439mbp

Sterman, J. D. (2002). All models are wrong: reflections on becoming a systems scientist. *System Dynamics Review*, *18*(4), 501–531. https://dx.doi.org/10.1002/sdr.261

Sweeney, L. B., & Sterman, J. D. (2000). Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review*, *16*(4), 249–286. https://dx.doi.org/10.1002/sdr.198

Woodworth, R. S., & Thorndike, E. L. (1901). The influence of improvement in one mental function on the efficiency of other functions. *Psychological Review*, *8*(3), 247–261. https://dx.doi.org/10.1037/h0074898

Venables, L., & Fairclough, S. H. (2009). The influence of performance feedback on goal-setting and mental effort regulation. *Motivation and Emotion*, *33*(1), 63–74. https://dx.doi.org/10.1007/s11031-008-9116-y

von Bertalanffy, L. (1968). *General system theory: Foundations, development, applications.* New York, NY: George Braziller.

Youmans, R. J., Figueroa, I. J., & Kramarova, O. (2011). Reactive task-set switching ability, not working memory capacity, predicts change blindness sensitivity. *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting*, *55*(1), pp. 914–918. https://dx.doi.org/10.1177/1071181311551190

Zaccaro, S. J. (2001). *The nature of executive leadership: A conceptual and empirical analysis of success.* Washington, DC: American Psychological Association. Available from http://dx.doi.org/10.1037/10398-000

Zaccaro, S. J., Weis, E. J., Chen, T. R., & Matthews, M. D. (2014). Situational load and personal attributes: Implications for cognitive readiness, adaptive readiness, and training. In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Teaching and measuring cognitive readiness* (pp. 93–115). Boston, MA: Springer. https://dx.doi.org/10.1007/978-1-4614-7579-8_5

Zimbardo, P. G., & Boyd, J. N. (1999). Putting time in perspective: A valid, reliable individual-differences metric. *Journal of Personality and Social Psychology*, *77*(6), 1271–1288. http://dx.doi.org/10.1037/0022-3514.77.6.1271

Appendix A: Comparison With Other Measures of Systems Thinking

Table A-1

How MATB Features Map Onto the Systems Thinking Components

MATB	Components
• Pumps turn on/off and break randomly • Using automation at the right time • Knowing when certain tasks are de-emphasized MATB is a “stock” and each task’s contribution to the total score is a “flow”	• Forecasting • Adaptive and flexible thinking
• Interdependencies in resource management task • Timing automation	• Closed-loop thinking • Forecasting
• Changing fuel levels in resource management • Valves on/off and reserve tanks empty/not empty • Turbulence (temporal changes) MATB is a “stock” and each task’s contribution to the total score is a “flow”	• Closed-loop thinking • Forecasting • Adaptive/flexible thinking
• Automation • Interdependencies in resource management	• Closed-loop thinking • Forecasting
• Resource management	• Closed-loop thinking • Holistic thinking

Appendix B: Construct Definitions and Experimental Hypotheses

ST Final Definitions

Systems thinking: The capacity to understand and capitalize on the interconnectedness of individual components of a system, both in the short-term and in the long-term.
Holistic thinking: The capacity to perceive a system as being greater than the sum of its parts, and to recognize how each component contributes (both directly and indirectly) to the output of the whole.
Closed-loop thinking: The capacity to detect causal relationships and feedback loops between and within components of a system.
Forecasting: The capacity to extrapolate current system state information to inform an understanding of likely future system inputs and outputs.
Adaptive and flexible thinking: The capacity to maintain awareness of holistic and causal relationships in the face of changing rules and/or goals and to alter plans and projections in light of changing information.

Hypotheses

Holistic Thinking

Positive relationship:

- **Analysis-holism scale (holistic thinking tendency)** will have the strongest positive relationship with holistic thinking, because it is essentially measuring holistic thinking.

- **Gestalt completion,** which measures speed of closure**,** will have a strong positive relationship with holistic thinking because the purpose is to combine pieces into one complete pattern.

- The word series task, which measures **inductive reasoning and fluid intelligence,** will have a strong positive relationship with holistic thinking because it is about pulling pieces together into one picture.

- **Cognitive complexity** will be moderately positively related to holistic thinking, because it considers parts of a system but will be more closely related to closed-loop thinking.

- **Pattern recognition** will be moderately positively related to holistic thinking because it involves understanding the meaning of patterns, which requires understanding a system as a whole.

- **Visualization** will be moderately positively related to holistic thinking because it is similar to speed of closure in the Gestalt task and requires an understanding of the whole.

- **Situation awareness** will have a positive relationship with ST because it considers an entire system in time and space.

Negative relationship:

Analytic thinking will be negatively related to holistic thinking, because it is the opposite — thinking of a system in terms of its parts instead of as a whole.

Unrelated:

- **Time perspective** will be unrelated to holistic thinking and will be closer related to closed-loop thinking.

- **Adaptive thinking/adaptive expertise** will be unrelated to holistic thinking. It will be most related to adaptive/flexible thinking and more related to closed-loop thinking or forecasting than holistic thinking.

- **Cognitive flexibility** will be unrelated to holistic thinking because it is about flexible thinking and does not require seeing a system as a whole.

- **Working memory** will be unrelated to holistic thinking.

Closed-Loop Thinking

Positive relationship

- **Pattern recognition** will be strongly positively related to closed-loop thinking because it requires seeing relationships between parts of a system.

- **Inductive reasoning** will be strongly positively related to closed-loop thinking because it requires seeing patterns and relationships.

- **Visualization** will be moderately positively related to closed-loop thinking because it requires seeing relationships between parts of a system when they are moved.

- **Cognitive complexity** will be moderately positively related to closed-loop thinking because it requires understanding cause and effect. However, it will be more related to forecasting because it regards predicting the future.

- **Analytic thinking** will be moderately positively related to closed-loop thinking because it involves breaking down a system into its parts.

- **Situation awareness** will be positively related to closed-loop thinking because it involves changing status in a system.

- **Holistic thinking** will be positively related to closed-loop thinking because seeing *all* cause and effect relationships within a system requires the ability to consider the entire system.

- **Time perspective** will be weakly positively related to closed-loop thinking because there is a temporal component to understanding cause and effect.

- **Cognitive flexibility** will be weakly positively related to closed-loop thinking because there is a component of change in cause and effect.

- **Working-memory** will be weakly positively related to closed-loop thinking because it requires only looking at the important information while blocking everything else out, which is necessary for understanding cause and effect.

Unrelated

- **Speed of closure** will be unrelated to closed-loop thinking because closed-link thinking does not in itself require speed.

- **Adaptive thinking** will be unrelated to closed-loop thinking because it requires coming up with an action in response to a system, not understanding effects in a system.

Forecasting

Strong positive relationship

- **Inductive reasoning** should be strongly and positively related to forecasting. Individuals high in inductive reasoning ability can identify the rules of the tasks quickly. Based on these rules, they should be able to predict the changing pattern of the task and prepare to take actions beforehand.

- **Future time perspective** should be strongly and positively related to forecasting. Individuals high in future time perspective care more about the future than other people. They are more likely to consider problems that are likely to happen in the future; therefore, they are more likely to utilize information they currently have to predict future events.

- **Cognitive complexity** should be strongly and positively related to forecasting. Individuals higher in cognitive complexity tend to integrate information about a larger number of individual components and use that information to make conclusions and predictions about events around them. Therefore, they should be better at predicting future events because they can better understand the interconnections among task components.

- **Speed of closure** should be strongly and positively related to forecasting. Individuals high in speed of closure can combine and organize individual pieces of information and identify meaningful patterns of events from the information. In an ST context, they should be better at integrating information about details to understand how things are likely to proceed; therefore, they should be able to make more accurate predictions about the future.

Moderate positive relationship

- **Pattern recognition** should be moderately and positively related to forecasting. Similar to speed of closure, individuals high in pattern recognition ability are better at identifying meaningful patterns; therefore, it should be positively related to forecasting. However, in contrast to speed of closure, pattern recognition focuses more on spatial-relevant ability and focuses less on dynamic events. Therefore, its relationship with forecasting should be weaker than speed of closure.

- **Working memory** should be moderately/strongly and positively related to forecasting. Individuals high in working memory can process a larger amount of information at the same time. Working memory can amplify the effects of cognitive complexity. Cognitive complexity increases the tendency to understand the events by analyzing a larger amount of information, and larger working memory increases the maximum amount of information that can be analyzed. This is a moderation rather than direct effect, so it may not be as strong; but considering that working memory could also relate to inductive reasoning and speed of closure, the relationship could also be very strong. Therefore, working memory could be moderately/strongly positively related to forecasting.

- **Situation awareness** should be moderately and positively related to forecasting. Individuals high in situation awareness can collect more information about current events than others. This information can be used to understand how events may change and to make predictions for the future. Situation awareness is also likely to amplify the effects of other variables that influence forecasting by facilitating the understanding of current/individual pieces of information (i.e., inductive reasoning, cognitive complexity, and speed of closure). With a larger amount of information, the accuracy of predictions made through inductive reasoning and complex thinking should be more accurate. Therefore, situation awareness should be positively related to forecasting, but the relationship may not be very strong because it is a moderation effect rather than a direct effect.

Weak relationship or unrelated

- **Holistic thinking** could be slightly and positively related to forecasting or it could be unrelated. Considering that the tendency toward thinking about the whole may be positively related to speed of closure, there may be a weak positive relationship between holistic thinking and forecasting. Other than that, there is no obvious relationship between holistic thinking and forecasting.

- **Analytic thinking** could be slightly and negatively related to or unrelated to forecasting. Considering that analytic thinking is the opposite end of holistic thinking, there may be a weak negative relationship with forecasting. Other than that, there is no obvious relationship between analytic thinking and forecasting.

- **Cognitive flexibilit**y should be unrelated to forecasting.

- **Adaptive thinking** should be unrelated to forecasting.

- **Visualization** should be unrelated to forecasting.

Adaptive and Flexible Thinking

Strong positive relationship

- **Cognitive flexibility** is the "acknowledgment of possible adjustments based on situational factors" (Martin & Rubin, 1995). It should be highly positively related to adaptive and flexible thinking, because adaptive and flexible thinking involve the ability to alter cognitions in the face of changing information, which is based on acknowledging possible adjustments.

- **Adaptive thinking** is the ability to "critique a situation in order to identify potential problems...and generate a set of alternative actions" (Joung, Hesketh, & Neal, 2006, p. 283). It should be highly positively related to adaptive and flexible thinking, because identifying potential problems and generating alternative solutions may result from an awareness of holistic and causal relationships in the system and may also lead to maintenance of awareness of these relationships.

- **Working memory** has been defined as the ability to block out irrelevant information and control attention (cf. Engle, 2002). It should be highly positively related to adaptive and flexible thinking, because it may facilitate maintaining awareness of important relationships in the face of changing information, by allowing a person to focus on the relevant information and not be distracted by losing or gaining superfluous information.

- **Pattern recognition** is "the process [in] which a person distinguishes a pattern he percepts with others and identifies what it means" (Pi, Lu, Liu, & Liao, 2008). It should be highly positively related to adaptive and flexible thinking, because recognizing patterns is part of recognizing and understanding holistic and causal relationships. It may also help with recognizing the signs of change and understanding what fundamental changes will be most impactful, and adjustments they might require.

Somewhat positive relationship

- **Visualization** is the "ability to imagine how something will look when it is moved around or when its parts are moved or rearranged" (Fleishman, Buffardi, Allen, & Gaskins, 1990). It may be somewhat positively related to adaptive and flexible thinking because imagining rearranged appearances might help with maintaining awareness of relationships in a visually-represented system. However, many systems do not have relationships that are visually-based, so the relationship is not likely to be strong.

- "**Analytic thinking** involves understanding a system by thinking about its parts" (Nisbett, Peng, Choi, & Norenzayan, 2001). It has been defined as "the tendency to understand the behavior of an object in terms of cause-and-effect, applying logic and reasoning to predict and explain outcomes" (Nisbett et al., 2001). It should be somewhat positively related to adaptive and flexible thinking because it involves the ability to predict cause and effect, which will help with adapting to changes in the system. It is also likely inversely related to having a holistic view of the system and seeing relationships, so the correlation is probably not extremely high.

- **Cognitive complexity** is "how individuals understand and predict the events happening around them," and a tendency to understand constructs using a greater number of individual components (i.e., un-simplified). In other words, greater cognitive complexity is associated with more highly differentiated depictions of constructs in memory (Bieri, 1955). It may be somewhat positively related to adaptive and flexible thinking because understanding and predicting events may lead to better adjustments. However, it is possible that a person could understand a system by seeing it holistically, rather than understanding the individual components, so the relationship may not be extremely high.

- **Inductive reasoning** is the "ability to combine separate pieces of information...to form general rules or conclusions" (Fleishman, Quaintance, & Broedling, 1984). It should be somewhat positively correlated with adaptive and flexible thinking, because combining new information with what already exists in the system may help with forming new solutions. However, this doesn't appear to be related to understanding the relationships in the system, so the relationship may not be extremely high.

- **Time perspective** is a psychological perception that orients people towards either the past, present, or future. It is a big-picture perspective: the tendency to understand present events in terms of how they might eventually impact future ones (Zimbardo & Boyd, 1999). It may be somewhat positively related to adaptive and flexible thinking, because it might help with understanding causal relationships and predicting change, but it does not provide any help with generating new solutions, so the relationship may not be extremely high.

- **Situation awareness** is "the perception of environmental elements within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995a). It may be somewhat positively related to adaptive and flexible thinking, because perceiving elements may lead to an earlier or more complex-

recognition and understanding of change. It might also lead to a better understanding of existing relationships within the system. However, it does not imply the ability to generate alternative solutions in the changing environment, so the relationship may not be extremely high.

Unrelated

- **Speed of closure** is the "degree to which different pieces of information can be combined and organized into one meaningful pattern quickly" (Fleishman, Quaintance, & Broedling, 1984). We are not expecting a significant correlation with adaptive and flexible thinking. Although combining information into a meaningful pattern can help with seeing relationships within a system, and doing so quickly may help with the integration of new information, one must also recognize the change, alter his or her cognitions in response, and react in order to be employing adaptive and flexible thinking. Speed of closure has very little to do with the totality of what is required here.

- **Holistic thinking tendency** is the tendency to think that every element of a task may be interconnected and cannot be understood in isolation from the whole (Choi, Dalal, Kim-Prieto, & Park, 2003). While this may help with understanding existing relationships within a system, it has nothing to do with the ability to recognize new information and react to it, which is required for adaptive and flexible thinking, so we are not expecting a significant correlation here.

- **Appendix C: Pilot Test Results**

Preliminary Analyses

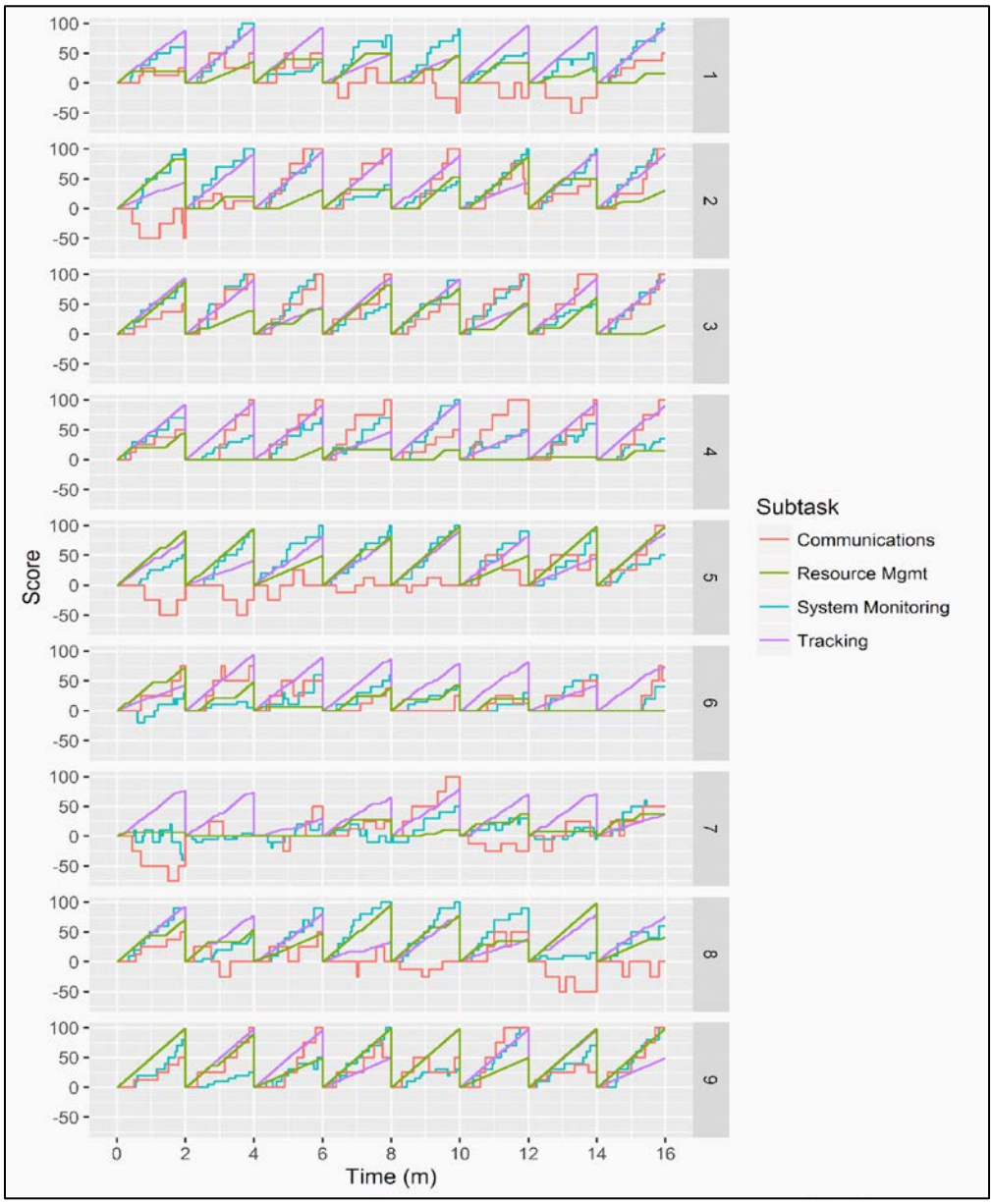

Figure C-1. MATB analyses of first nine participants (May 31, 2016). Represents change in score over time. This graph depicts the point changes over time for each participant, split by MATB subtask. Tracking seems to be the most consistent source of points, while communications seems to be the most variable (frequently going negative).

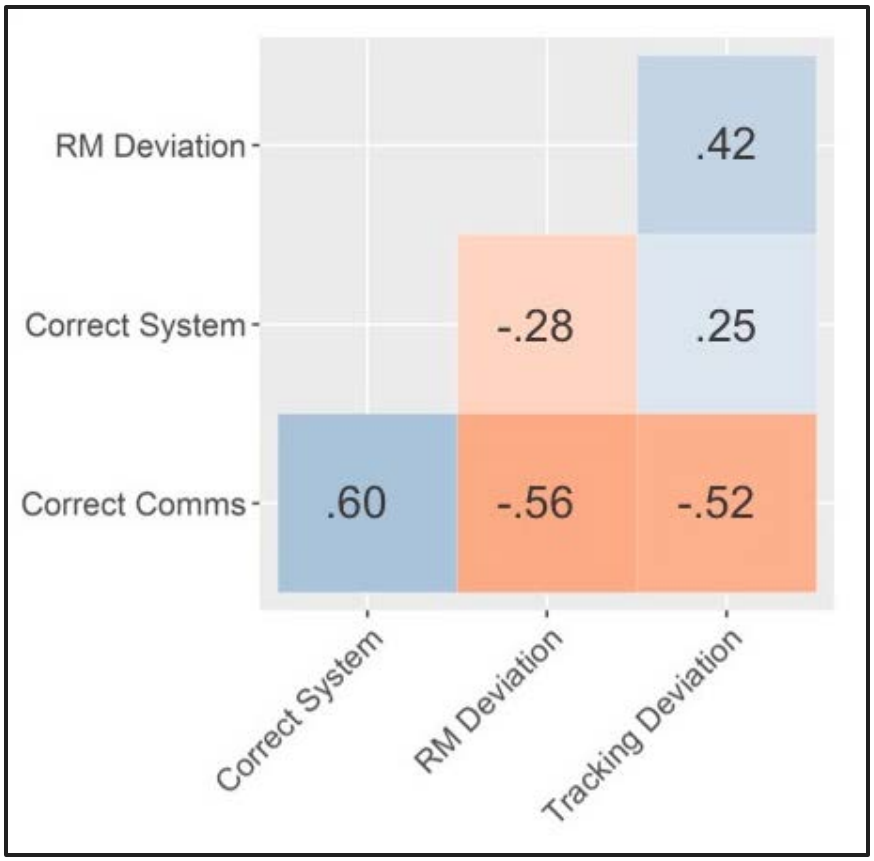

Figure C-2. Correlation matrix, subtasks (performance aggregated over epochs). All values are non-significant. Correlations between subtasks are moderate-to-high. Lack of significance is likely attributable to small sample size. Note that Comms = Communications, and RM = Resource Management.

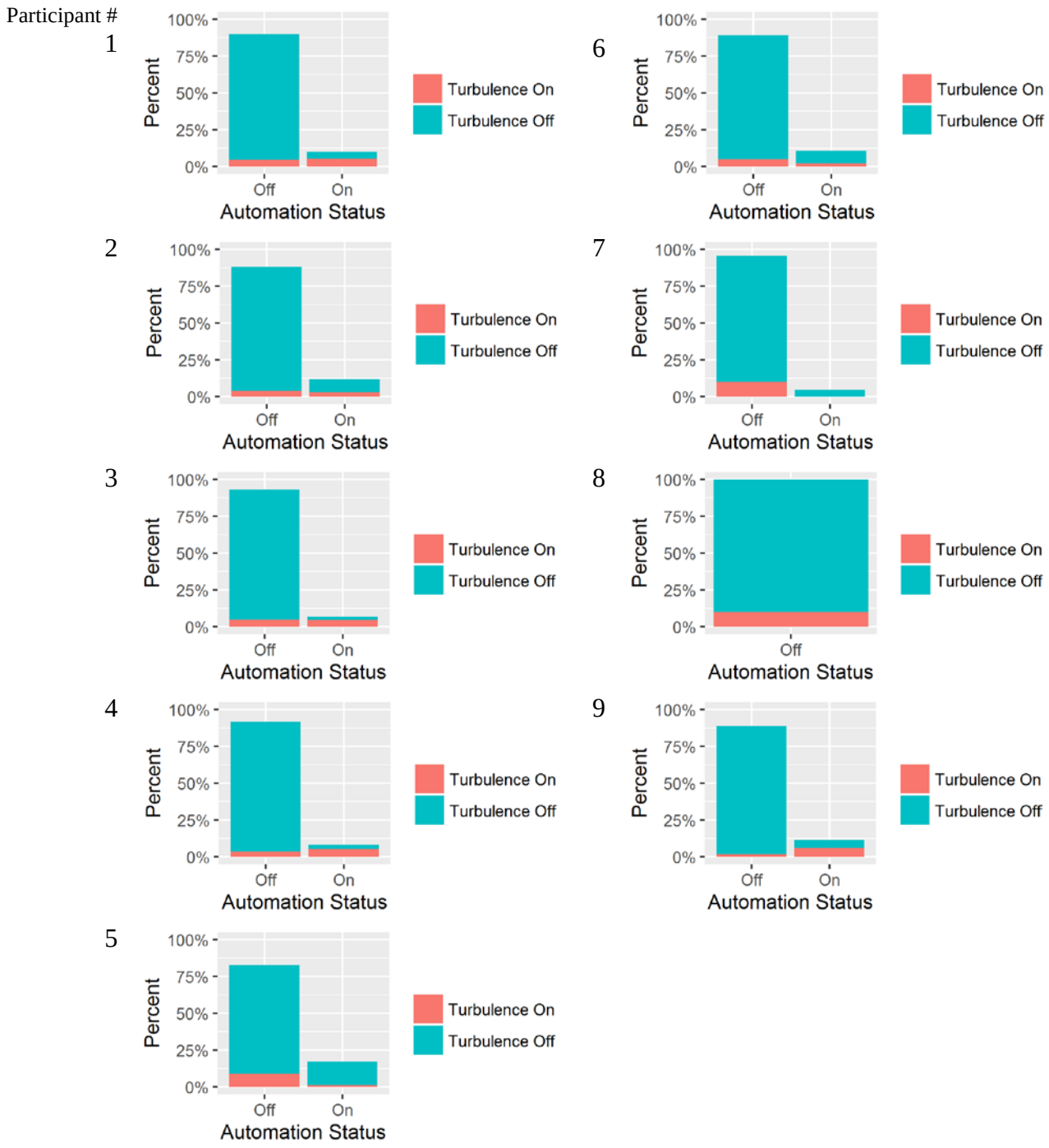

Figure C-3. Data for each pilot participant showing the overlap of turbulent events with engagement of automation. The blue shading indicates when there was no turbulence and the automation was technically not required, whereas the red shading shows when turbulence was on and automation was required. The *x*-axis displays the action taken by the participant; in other words, whether the automation was turned "on" or "off." The *y*-axis in each figure represents the

percent of time automation was engaged and not engaged during a turbulence event. To further explain, consider the first figure in the sequence. This graph shows that when automation was "on" (right side of figure), the participant was able to overlap a considerable portion of the turbulence event, indicated by the red shading. When the automation was "off," meaning the participant took no action (left side of the figure), it can be seen that there was a portion of time when the participant should have engaged automation but didn't (indicated by the red) and a much larger portion when the participant didn't engage the automation and wasn't required to (indicated by the blue).

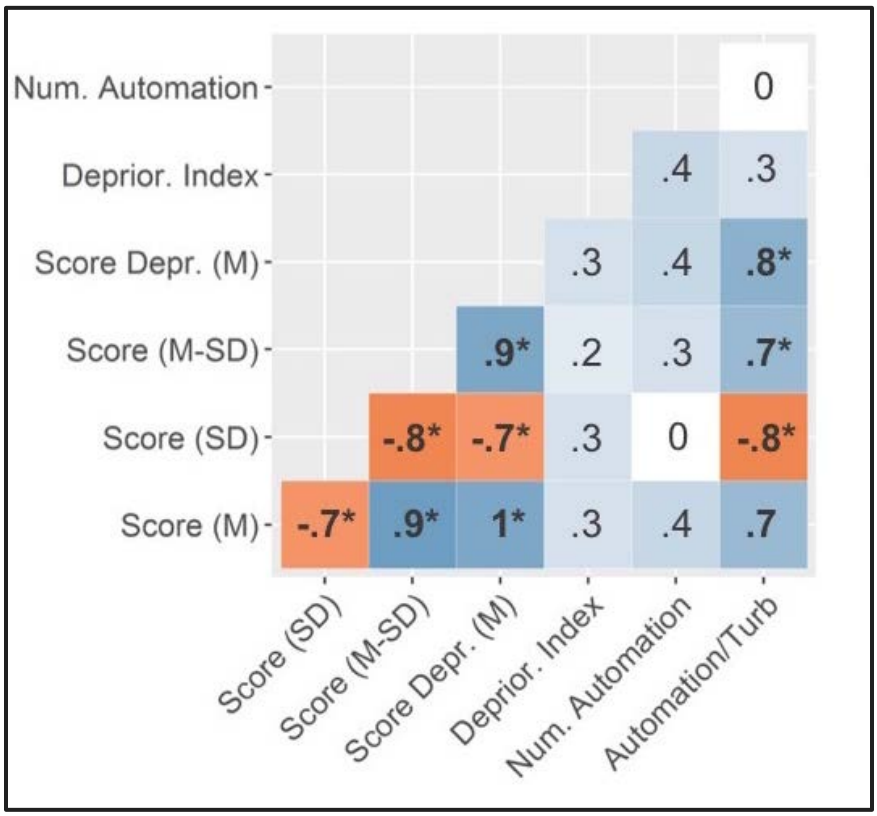

Figure C-4. Closed loop thinking correlation matrix. Correlations are rounded to the nearest tenth; * indicates significant values ($p > .05$). Deprior Index = Deprioritization Index.

Limitations

At the current stage of development of the ST model, there are a few ideas about additional changes that could be made to the model in the future. First, there is a need to brainstorm closed-loop thinking further. Currently there are only a few, if any, dependencies for the closed-loop thinking measure. Secondly, pumps breaking is not currently recorded. It would be useful to have this information for the adaptive/flexible thinking score gain post-break. Last, it might be useful to record the "reserve tank" capacity for use as a measure for forecasting and/or holistic thinking. The only reason to keep the reserve filled is to modulate the speed of refill down the line.

Appendix D: ST Criterion Validation Research Study Tables and Figures

Table D-1

Means, Standard Deviations, Reliability, and Correlations

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Future time perspective	4.09	.68	(.73)																	
2. Holistic thinking tendency	4.87	.57	.34	(.76)																
3. Cognitive complexity	4.25	.67	.34	.37	(.91)															
4. Cognitive flexibility	4.58	.82	.34	.23	.31	(.90)														
5. Situation awareness	2.65	.54	.42	.18	.14	.18	(.81)													
6. Adaptive thinking	4.43	.60	.10	.25	-.04	-.11	.13	(.86)												
7. Speed of closure	12.88	2.99	.12	.04	.17	.16	.00	-.25	NA											
8. Visualization	5.01	2.66	.00	.12	.04	.09	-.08	-.09	.30	NA										
9. Pattern recognition	12.17	15.69	-.03	.06	.04	.06	-.04	-.07	.26	.58	NA									
10. Word series test	18.22	6.07	.08	.14	.10	.21	-.10	-.07	.19	.39	.41	NA								
11. Letter series test	36.71	8.71	.01	.06	-.03	.02	-.05	.05	.01	-.04	.01	.14	NA							
12. Analytic thinking	1.73	1.24	.01	-.01	.04	.02	.02	-.05	.11	.27	.28	.24	-.06	NA						
13. Numeracy	11.22	2.44	.12	.22	.16	.12	.00	-.01	.23	.42	.34	.36	-.05	.31	NA					
14. STQ	76.74	10.71	.30	.42	.53	.34	.00	-.19	.23	.26	.18	.32	-.01	.14	.35	(.76)				
15. ST task	1.78	.89	.03	.10	.05	.05	-.01	-.13	.28	.36	.31	.25	-.05	.26	.36	.19	NA			
16. Job performance	5.31	1.09	.26	.17	.10	.27	.19	.14	-.06	.01	-.02	.15	.03	-.01	.06	.17	-.02	NA		
17. Received praise	18.87	2.83	.30	.26	.27	.23	.17	-.01	.11	.09	.00	.13	.05	.03	.12	.26	.07	.35	NA	
18. Job complexity	.00	2.10	.16	.04	.13	.08	.25	.15	-.18	-.11	-.09	-.05	-.02	-.02	-.12	.01	-.06	.16	.16	NA

Note. N = 406. Coefficient alphas appear on the diagonal. NA indicates where reliability is not applicable for competency measures or computed variables. STQ = 15-item ST Questionnaire (Davis & Stroink, 2015) in which ST is conceptualized more as a dispositional tendency/preference. ST Task = 4 graph-reading questions (Sternman, 2002) measuring basic ST concepts (e.g., stocks and flows, time delays), in which ST is conceptualized more as an ability or skill.

Table D-2

Hierarchical Regression Results for Incremental Validity and Moderation Effects Involving STQ

	Job performance		Received praise	
Antecedent variable	Step 1	Step 2	Step 1	Step 2
Future time perspective	.26**	.23**	.29**	.25**
Numeracy	.03	.00	.09	.03
STQ		.10		.18**
R^2	.07	.08	.10	.13
ΔR^2		.01		.03**
Holistic thinking tendency	.17**	.13*	.25**	.18**
Numeracy	.02	.02	.07	.02
STQ		.12*		.18**
R^2	.03	.04	.07	.10
ΔR^2		.01*		.02**
Cognitive complexity	.09	.02	.25**	.18**
Numeracy	.05	.00	.08	.04
STQ		.16*		.15*
R^2	.01	.03	.08	.09
ΔR^2		.02*		.02*
Cognitive flexibility	.26**	.24**	.22**	.15**
Numeracy	.03	.00	.10*	.04
STQ		.09		.20**
R^2	.07	.08	.06	.09
ΔR^2		.01		.03**
Situation awareness	.19**	.19**	.17**	.17**
Numeracy	.06	.00	.12*	.03
STQ		.17**		.25**
R^2	.04	.07	.04	.10
ΔR^2		.02**		.06**
Adaptive thinking	.14**	.18**	-.01	.04
Numeracy	.06	-.01	.12*	.03
STQ		.20**		.26**
R^2	.02	.06	.02	.07
ΔR^2		.04**		.06**
Speed of closure	-.08	-.11*	.08	.05
Numeracy	.08	.02	.10*	.03

STQ		.18**		.24**
R^2	.01	.04	.02	.07
ΔR^2		.03**		.05**
Visualization	-.01	-.04	.05	.02
Numeracy	.07	.01	.10	.03
STQ		.17**		.25**
R^2	.00	.03	.02	.07
ΔR^2		.03**		.05**
Pattern recognition	-.05	-.06	-.04	-.06
Numeracy	.08	.02	.14**	.05
STQ		.17**		.26**
R^2	.01	.03	.02	.07
ΔR^2		.03**		.06**
Word series test for inductive reasoning	.15**	.12*	.10	.05
Numeracy	.01	-.03	.09	.02
STQ		.14**		.24**
R^2	.02	.04	.02	.07
ΔR^2		.02**		.05**
Letter series test for inductive reasoning	.04	.03	.05	.05
Numeracy	.06	.00	.12*	.03
STQ		.18**		.26**
R^2	.01	.03	.02	.08
ΔR^2		.03**		.06**
Analytic thinking	-.03	-.03	-.01	-.02
Numeracy	.07	.01	.13*	.04
STQ		.17**		.25**
R^2	.00	.03	.02	.07
ΔR^2		.02**		.06**
STQ	.16**		.24**	
Job complexity	.16**		.16**	
Numeracy	.02		.05	
STQ×Job complexity	-.01		.07	
R^2	.05		.10	
Adjusted R^2	.04		.09	

Note. N = 406. *p < .05. **p < .01. STQ = 15-item ST Questionnaire (Davis & Stroink, 2015) in which ST is conceptualized more as a dispositional tendency/preference.

Table D-3

Direct and Indirect Effects Mediated by STQ

Outcome	Job performance			Received praise		
Antecedent	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
Future time perspective	.37**	.00	.37**	1.03**	.06**	1.09**
Holistic thinking tendency	.23*	.00	.24*	.86**	.06**	.92**
Cognitive complexity	.03	.00	.03	.76**	.04**	.80**
Cognitive flexibility	.32**	.00	.32**	.51**	.04**	.55**
Situation awareness	.36**	.01**	.37**	.88**	.06**	.94**
Adaptive thinking	.33**	.01**	.34**	.14	-.01	.14
Speed of closure	-.04*	.00	-.04*	.04	.00	.04
Visualization	-.01	.00	-.01	.02	.00	.02
Pattern recognition	.00	.00	.00	-.01	.00	-.01
Word series test for inductive reasoning	.02**	.0003**	.02**	.02	.00	.03
Letter series test for inductive reasoning	.00	.00	.00	.02	.00	.02
Analytic thinking	-.03	.00	-.03	-.06	.00	-.06

Note. N = 406. Numeracy was included as a control variable in all analyses. *p < .05. **p < .01. STQ = 15-item ST Questionnaire (Davis & Stroink, 2015) in which ST is conceptualized more as a dispositional tendency/preference

Table D-4

Direct and Indirect Effects Mediated by ST Task

Outcome	Job Performance			Received Praise		
Antecedent	Direct Effect	Indirect Effect	Total Effect	Direct Effect	Indirect Effect	Total Effect
Future time perspective	.41**	-.02	.39**	1.22**	.14	1.36**
Holistic thinking tendency	.33**	-.02	.31**	1.24**	.09	1.33**
Cognitive complexity	.14*	-.01	.14	1.05**	.09	1.14**
Cognitive flexibility	.36**	-.02	.34**	.71**	.05	.76**
Situation awareness	.39**	-.02	.36**	.84**	.11	.94**
Adaptive thinking	.24*	-.01	.24*	-.06	-.01	-.07
Speed of closure	-.03	.00	-.03	.09	.00	.09
Visualization	.00	.00	.00	.05	.01	.06
Pattern recognition	.00	.00	.00	-.01	.00	-.01
Word series test for inductive reasoning	.03**	.00	.03**	.05	.00	.05
Letter series test for inductive reasoning	.00	.00	.00	.02	.00	.02
Analytic thinking	-.02	.00	-.02	-.05	.00	-.05

Note. N = 406. Numeracy was included as a control variable in all analyses. *p < .05. **p < .01. ST Task = 4 graph-reading questions (Sternman, 2002) measuring basic ST concepts (e.g., stocks and flows, time delays), in which ST is conceptualized more as an ability or skill.