Received: 9 May 2023 | Revised: 15 February 2024 | Accepted: 31 May 2024

DOI: 10.1002/sres.3033

RESEARCH PAPER

SYSTEMS and BEHAVIORAL RESEARCH SCIENCE WILEY

A multi-dimensional index of evaluating systems thinking skills from textual data

Ning-Yuan Georgia Liu[1,2] | **Hesam Mahmoudi**[1,2] | **Konstantinos Triantis**[1] | **Navid Ghaffarzadegan**[1]

[1]Grado Department of Industrial and Systems Engineering, Virginia Tech, Falls Church, Virginia, USA

[2]MGH Institute for Technology Assessment, Harvard Medical School, Somerville, MA, USA

Correspondence
Ning-Yuan Georgia Liu, MGH Institute for Technology Assessment, Harvard Medical School, 399 Revolution Drive STE 1190, Somerville, MA, 02145, USA
Email: ningyuan@vt.edu; gliu26@mgh.harvard.edu

Funding information
national science foundation, Grant/Award Number: 1824594

Abstract

Systems thinking (ST) includes a set of critical skills and approaches for addressing today's complex societal problems. Therefore, it has been introduced into the curricula of many educational programmes around the world. Despite all the attention to ST, there is less consensus when it comes to evaluating and assessing ST skills. Particularly, a quantitative assessment approach that captures ST's multi-dimensionality is crucial when evaluating the degree to which one has learned and utilizes ST. This paper proposes a systematic approach to create such a multi-dimensional Index of ST from textual data. Initially, we provide an overview of the theoretical background as it pertains to different measurement approaches of ST skills. Then we provide a conceptual framework based on ST skill measures and transform this framework into a quantifiable model. Finally, using student data, we provide an illustration of an integrated index of ST skills. We compute this index by using a mixed methods approach, including robust principal component analysis, data envelopment analysis and two-staged bootstrapping approach. The results show that (i) our model serves as a systematic multi-dimensional ST approach by including multiple measures of ST skills and (ii) international students and self-reported math skills are found as significant predictors of one's level of ST in the graduate student dataset ($N = 30$), however no significant factors are found in the first-year engineering student dataset ($N = 144$).

KEYWORDS

data envelopment analysis, engineering education, mental maps, problem structuring, systems thinking

1 | INTRODUCTION

Systems thinking (ST) is the ability to see the world as complex systems and recognize that its components are interconnected, influence each other, often, in unexpected ways and as such are hard to intuit (Cabrera & Cabrera, 2019; Sterman, 2000; Wolstenholme, 1993). With the growth in the complexity of societal and

technological problems, increasing attention is being paid to training individuals to recognize complex systems and improve ST skills (Fordyce, 1988; Nehdi & Rehan, 2007). However, among several challenges, less consensus exists around the assessment of ST skills (Dugan et al., 2022). The problem is partly related to various definitions of ST and its multi-dimensionality, which make it hard to assess by a single measure (Mahmoudi et al., 2019). The literature offers a range of ST measurement methods varying from self-assessment surveys (Huz et al., 1997), verbal protocol analysis (Maani & Maharaj, 2004) and scenarios for evaluating the understanding of complexity (Sweeney & Sterman, 2000), including simulation games, and computer-based or board-based role-playing that assess one's understanding of feedback loops and system delays (Barlas & Diker, 2000; Kunc & Morecroft, 2007; Lane, 1995; Sterman, 1989a, 1989b).

There have been reviews of the ST measurement literature, identifying drawbacks of the measurement methods or inviting improvements of the measurement techniques. For example, the methods relying on self-assessment are known to suffer from biases commonly attributed with self-reported measures (Davis et al., 2020, 2023; Cavaleri & Sterman, 1997) since individuals are often unaware of their own biases and misconceptions of systems (Hahn & Gawronski, 2019). On the other end of the spectrum, assessment methods focusing on behaviour or patterns of thought are resource-consuming and difficult to administer in large numbers (Maani & Maharaj, 2004). More importantly, it is argued that most of the available assessment methods focus on one or few skills out of a possible set of ST skills (Plate, 2010). With respect to simulation games, concerns about external validity and representativeness are often raised questioning the extent to which the game context represents the real world (Hammond & Stewart, 2001; Samoylova, 2014). Even ST assessment techniques that try to cover a wide range of skill sets, often, focus on a single measure or simply add up different measures without fully acknowledging the multi-dimensionality of ST skill sets.

To expand on the multi-dimensionality of ST assessment, it is important to note that most ST definitions represent ST as a skill set, that is, demonstrating high levels of ST indicates performing well on a set of ST-related tasks and demonstrating a set of skills (Richmond, 2000). The ST skill sets that are commonly used as a basis for assessment are represented as different levels of sequential characteristics, similar to how people learn as they move up in a schooling system, from basic concepts to more sophisticated concepts (Mahmoudi et al., 2019). The fact that the majority of ST definitions use different levels of sequential ST characteristics highlights the relevance and importance of multi-dimensional assessment approaches.

The objective of this paper is to compute a multi-dimensional ST index where ST skill characteristics are considered concurrently. While there are several valuable perspectives that can offer differing definitions of ST (Jackson, 2016), our perspective is closer to the system dynamics school of thought that emphasizes the role of seeing the world as an interconnected system that comprises delays and reciprocal causality in the form of feedback loops (Forrester, 1997; Sterman, 2000). Our specific contributions in this paper are twofold: First, we offer a correspondence of mental map measures to ST skills. Second, we propose a systematic and rigorous approach for computing a multi-dimensional index of ST based on measured ST skills. We implement the method to two datasets, namely, the first of 144 undergraduate students and the second of 30 engineering graduate students who responded to a realistic, scenario-based, open-ended question that is concerned with a complex socio-environmental problem. A novelty of this paper is the mapping efficiency performance concepts (Charnes et al., 1994) to the literature of ST and ST skill set assessment. Our proposed index can be used for educators and stakeholders as a framework for assessing ST skills. The Lake Urmia Vignette (LUV) (Davis, et al. 2020) serves as an example of how the proposed index can be implemented. The index is not limited to the LUV, but it can be applied to any type of case studies and vignette that depicts a complex problem. The rest of the paper is organized as follows: Section 2 provides a brief theoretical background of our research. Section 3 presents our methods and dataset. Section 4 presents our modelling using the multi-dimensional optimization approach of data envelopment analysis (DEA). Section 5 discusses the results obtained from our analysis. Finally, Section 6 summarizes the key findings, and Section 7 concludes with future research.

2 | THEORETICAL BACKGROUND

Prior studies introduced different tools and approaches to measure ST skills and have investigated interventions that can help improve them. First, a common approach, especially in the education and psychology literature, includes self-assessment surveys. Researchers often provide a set of questions designed to ask the participants to assess some aspects of their decisions and perceptions to identify their ST skills. Commonly used surveys of ST are the critical openness survey and the reflective scepticism survey. It is argued that responses to these surveys are often within the same range and do not vary enough

for researchers to distinguish individuals (Davis et al., 2020). This might be due to the fact that such self-assessment surveys rely too much on one's perception of their skills, which can be affected by self-attribution or overconfidence biases or simply by lack of awareness about own skills (Davis et al., 2023; Pike, 2011; Porter, 2011).

The second group includes simulation games that measure one's decisions when facing a complex problem rather than one's perception of their own skills. Such games can be conducted in different formats including computer simulation games or board games (Kunc & Morecroft, 2007). Sterman's (1989a) study of people's decision in a production distribution game is a good example of the approach and how it can help evaluate peoples' understanding of feedback loops and system delays. The research resulted in coining the term, misperception of feedback, that is, the human failure to understand feedback loops. This concept was later supported by more studies. Howie et al. (2000) and Moxnes (2000, 2004) introduced experimental simulation games to assess dynamic decision-making skills and the understanding of feedback loops. Following this line of research, effects of delay, non-linearity and feedback have been investigated in stock management simulation games. For example, in Özgün and Barlas (2015), participants play the role of a production manager of a T-shirt production. The objective of the game is to bring the inventory level to a target level as quickly as possible. Participants are given a range of tasks associated with each production factor. Then their performance outcome and perceived difficulty ratings are measured.

The third group includes scenario-based assessments. In these tasks, participants are given a scenario about a problem, and their responses are evaluated based on some form of scoring rubric. One example is the department-store task, which is commonly used for assessing peoples' understanding of accumulation (Cronin et al., 2004). In this scenario-based assessment, study participants receive information about the rate at which people come to a store and leave the store and are asked to determine when the store reaches its maximum and minimum number of customers. The task has consistently shown peoples' misunderstanding of accumulation (Sweeney & Sterman, 2000), which has persisted even after interventions (Baghaei Lakeh & Ghaffarzadegan, 2015; Herrera-Restrepo et al., 2016; Hendijani, 2021). The approach is not limited to system dynamics. In Grohs et al. (2018) participants are given a short scenario and asked to respond to six questions related to the scenario. Their responses are analysed by exploring three different dimensions of ST: problem, perspective and time dimensions. These are evaluated by a specific scoring rubric. Similarly, the LUV, which is used in the current study, describes the problems associated with a shrinking lake in north-east of Iran. The scenario is then used toassess participants' understanding on the main causes of the shrinking lake (Davis et al., 2020).

Finally, and related to the third category of scenario-based assessments, at times, peoples' responses to a scenario are coded and turned to a graphical representation of their thought process, that is, a mental map (Mahmoudi et al., 2019; Montibeller & Belton, 2006; Richmond, 2000). Mental mapping is an intermediary step to describe mental models[1] and has traditionally been used as a source of information about system complexity (Kim, 2009; Kim & Andersen, 2012) or used as a communication tool (Black, 2013). Mental maps represent how one thinks about the world in terms of causal connections. The most common measures of ST skills that utilize mental maps examine the interconnectivity and complexity of the causal network using measures from network theory. For example, Levy et al.'s (2018) study grouped 149 cognitive maps into three main clusters with each distinct cluster representing the different combinations of common motifs. Other studies use semi-quantitative cognitive mapping techniques that capture network structure richness and web-like causality (Gray et al., 2019). These studies compare and contrast different network measures applied to mental maps (Haque et al., 2023; Naugle et al., 2021).

Even though the measurement tools vary, in the end, most assessment methods focus on one or few skills or outcome measures out of a range of possible ST skill sets. This limits the extent to which they represent one's comprehension of complexity. Notable exceptions are three efforts to assess different characteristics of ST skills. Table 1 depicts the sequence at which these three approaches view ST. First, Barry Richmond (1993) presents seven ST skills and argues that 'doing good systems thinking means operating on at least seven thinking tracks simultaneously' (p. 121). His definition suggests the multi-dimensional nature of ST by explicitly emphasizing the development of each skill separate from the rest of the skill set, that is, individuals can possess different levels of each of the seven ST skills. Second, Assaraf and Orion (2005) offers the ST hierarchy, basing it on eight, literature-supported, characteristics of ST. They place their eight ST skills on a sequence; yet, they addressed each of the skills separately in their assessment effort. Thus, their scoring systems and scores for each

[1]Forrester (1971, p. 112) defines a mental model as 'the mental image of the world around [us] that we carry in [our] heads', which are viewed as 'incomplete' and 'imprecisely stated', and constantly changing with time.

TABLE 1 Three well-accepted ST definitions with different approaches of considering ST skills.

Approach	Richmond (1993)	STH: Assaraf and Orion (2005)	STC: Stave and Hopper (2007)
Level 1	Specify problems: Forest thinking System as cause thinking Dynamic thinking	System components: The ability to identify the components of a system and processes within the system	Basic: Recognizing interconnections Identifying feedback Understanding dynamic behaviours
Level 2	Construct model: Quantitative thinking Closed loop thinking Operational thinking	Synthesis of system components: The ability to identify relationships among the system's components The ability to organize the systems' components and processes within a framework of relationships The ability to make generalization The ability to identify dynamic relationships within the system	Intermediate: Differentiations types of flow and variables Using conceptual models
Level 3	Test model: Scientific thinking	Implementation: Understanding the hidden dimensions of the system The ability to understand the cyclic nature of systems Thinking temporally: Retrospection and prediction	Advanced: Creating simulation models Testing policies

skill are separate from the others. Third, Stave and Hopper (2007) derive seven ST skills from experts and literature consensus and place them on their ST continuum. Then, they map the continuum to different levels of understanding, placing all seven skills on the same dimension. They also note that it is difficult to identify the exact order of the skills, hinting at the multi-dimensional nature of ST.

As discussed, these studies view ST as a continuous concept, like a ladder that one steps up to achieve full skills. Another major limitation of studies that acknowledge the different layers of ST skills is that in practice, the assessments are still based on a single ST measure or an arbitrarily weighted average of some of the ST measure (Davis et al., 2020; Yin et al., 2005). The approach that considers the weighted average of the ST skill measures can be justified based on expert opinion, but it is not systematic in the way it arrives at the weights. In sum, the literature does not offer an approach that considers the characteristics of ST skill sets concurrently that goes above and beyond additive measures or sequential approaches. This gap is the focus of our paper.

In this study, our interdisciplinary approach is at the intersection of operations research (OR) and ST (Ghaffarzadegan & Larson, 2018), and the novelty relates to the computation of an index by utilizing a non-parametric method—DEA—that is a powerful approach for multi-criteria performance assessment due to its ability to benchmark multi-dimensional inputs and outputs (Charnes et al., 1978; Cooper et al., 2011, 2014). This approach is applied in various contexts from operations management, maintenance and water infrastructures to healthcare systems and education (Andalib, 2018; Bhatkoti et al., 2018; Darabi et al., 2021). The innovation of our approach lies in the determination of the appropriate input and output ST measures based on the theoretical background discussed in this section and the generalizability of our method to any scenario where multiple ST skill measures need to be considered.

3 | METHOD

This paper utilizes a scenario-based task that elicits people's response to a complex socio-environmental problem. The task referred to as the LUV is a real-world case of a depleting natural lake with many economic, health and environmental consequences. Participants' answers to the broad question about the problem of the lake is in a text format, which we use to extract corresponding mental map and measures, as described in Section 3.1. The data are described in Section 3.2. Then we provide a modelling approach (Section 3.3) to consider all ST skill measures concurrently, based on the mental map measures associated with the ST skill set.

TABLE 2 Mental map measures and their definitions.

Measure	Definition	Source
Variables	Calculate the number of identified variables in the response. Represents a combination of detailed complexity, that is, identifying the components of a system, and dynamic thinking since we are focusing on variables rather than the components, which goes further than the identification of mere system components	Assaraf and Orion (2005); Richardson (1994); Stave and Hopper (2007)
Causal links	Calculate the number of causal links that connect identified variables. Represents a measure of interconnectivity, which resembles the notion of 'cause–effect thinking' and 'recognizing interconnections'	Dorani et al. (2015); Stave and Hopper (2007)
Closed loops	Calculate the number of closed loops in the response. This represents a measure of 'identifying feedback' as well as 'closed loop thinking'. Responses that contain closed loops in their causal network demonstrate the ability to view problems as ongoing dynamics of the loop rather than the result of an exogenous cause. In this sense, this measure is also related to the concept of 'endogenous viewpoint' and 'system-as-cause thinking'	Dorani et al. (2015); Richardson (1994, 2011); Stave and Hopper (2007)
Middle nodes	Calculated as the variables with an arrow going in and another arrow coming out of them. This measure represents the depth of the causal network. Responses whose causal networks contain higher middle nodes are those who contain loops, causal chains and intertwined structures, rather than isolated links and lists of causes or effects. This measure is in line with the skill of 'operational thinking' and 'endogenous viewpoint'	Haque et al. (2023)
Connectivity	Calculate by the number of identified variables subtracting the number of identified causal links.[a] This measure resembles 'link density',[b] which is a common measure of interconnectivity of causal networks and 'cyclomatic complexity',[c] which is suggested to represent the concept of interconnectivity	Naugle et al. (2021); Plate (2010)

[a]Since it is possible to have more variables than causal links, that is, one identified more variables than links, we make connectivity non-negative by adding the smallest number of variables minus causal links.
[b]Link density is calculated as the number of causal links divided by the number of variables (Plate, 2010).
[c]Cyclomatic complexity is calculated as number of causal links minus number of variables plus two times the number of connected components (Naugle et al., 2021).

3.1 | The scenario

The LUV task is a tested and validated method proposed by Davis et al. (2020) to assess individuals' understanding of complex systems. The LUV tool consists of a short case study about the real-world shrinkage of Lake Urmia due to various concerns. Participants are asked to read the vignette and explain in writing their understanding of what 'went wrong' in this complex system. The responses are then fed into a carefully designed rubric and coding procedure that transforms textual data into quantifiable measures of ST skills, including the number of variables, the number of causal links and the number of closed loops in the causal network. In addition to these measures, in our approach, we record the number of middle nodes as a fourth measure, that is, the number of times a variable is considered both as a consequence of an effect and a cause for another effect, schematically represented as A → middle node → B. Counting the number of middle nodes accounts for an individual's comprehension of a chain of causality. Moreover, we combine two of the mental map measures to isolate skills from the ST skill set and arrive at a measure of network connectivity. These variables provide the necessary input for the DEA formulations (Section 4.1). Table 2 summarize each of these five recorded mental map measures and how they are associated with in the ST skill set.

3.2 | Dataset description

We use a dataset collected from 144 from first-year engineering undergraduate students at Virginia Tech from their answers to the LUV ST assessment task (Davis et al., 2020). Out of 144 graduate student, 46 (32%) of them were female, and only 3 (2%) were international students (see descriptive statistics for the data in the Supporting Information: Appendix A, Table A1).

The variables collected in the dataset include (i) word count, which represents the numbers of words in a LUV response; (ii) the set of five ST measures, which are variables, links, loops, middle nodes and connectivity; (iii) demographic factors including age, gender, level of study (PhD/Masters) and nationality (international/US);

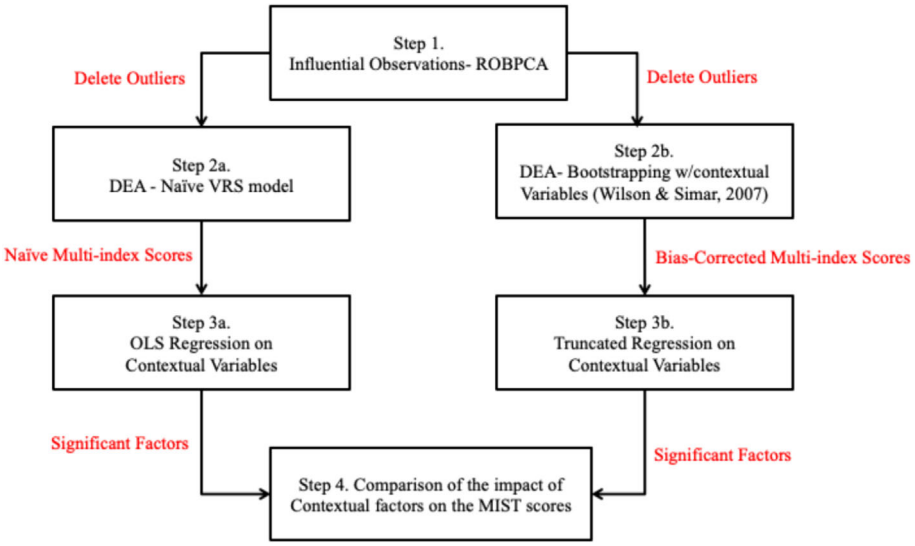

FIGURE 1 The modeling approach. [Colour figure can be viewed at wileyonlinelibrary.com]

and lastly, (iv) participant educational backgrounds, including their understanding of the basic concept of systems (e.g. accumulation, feedback loops, causal maps), self-reported math skills and familiarity with technical terms of systems (feedback loop questions).[2] The descriptive statistics for the dataset are provided in the Supporting Information: Table A1. In the Supporting Information: Appendix B, we test the replicability of the results using a different dataset collected from 30 graduate students recruited across two departments in the College of Engineering.

3.3 | The modelling approach

The overall framework of our approach is depicted by Figure 1. We start with an influential observation analysis using robust principal component analysis (ROBPCA; Hubert & Engelen, 2004) to identify groups of influential observations (step 1). ROBPCA is strong at distinguishing influential observations from regular observations, especially when analysing highly dimensional data. Using the distance measures, the method flags four types of observations: regular, good leverage, orthogonal and bad leverage points (Hubert & Engelen, 2004). We perform ROBPCA on three sets of data combinations to reveal the influential observations from different angles. First, we consider the input and output variables focusing solely on operational characteristics (Herrera-Restrepo et al., 2016). Second, we include all input, output and contextual variables by taking the contextual characteristics into account, and lastly, we consider only the contextual variables (Triantis et al., 2010). We then consider discarding the bad leverage points as outliers since they may indicate unusual behaviour in terms of the distance measures. Before discarding any observation, we complete a meta-analysis to make sure that an observation discarded is very different from the remainder of the dataset.

In step 2, we perform the DEA analysis with the model specifications of Section 4.1. There are two DEA models performed in this stage. Step 2a follows the traditional DEA performance measurement analysis, where one considers only the input and output system variables. The explanation of the model is provided in Section 4. In parallel, step 2b in addition to the input/output variables considers the contextual variables when computing performance scores. We conducted a two-staged bootstrapping technique, followed by a truncated regression on the contextual variables (Simar & Wilson, 2007). This technique introduces the bias-corrected performance scores and its confidence intervals so that it does not overestimate the performance scores as compared to traditional DEA analysis. We compare the naïve multi-index score from our approach (derived from step 2a) and the bias-corrected score (derived from step 2b) to assess the impact of the contextual variables. When using traditional DEA models (Section 4), a computed score may be biased if the contextual factors are not considered. This is relevant when measuring one's ST skills, because there are socio-demographic and educational background factors that could affect one's levels of ST (Naugle et al., 2021).

Finally, in step 3, we perform regression models to test if the contextual variables significantly predict the multi-index scores. Step 3a uses an ordinary least square (OLS) model by treating the naïve multi-index scores as the dependent variable and contextual variables as

[2]The detailed recruitment and data collection process, as well as further variable definitions, are documented in (Davis, 2020).

independent variables. Step 3b, performs a truncated regression by treating the bias-corrected multi-index score as the dependent variable, with the same sets of contextual variables. Finally, we identify the resulting significant contextual factors from the regression models in step 4.

4 | MODELLING: DEA

Our proposed index is computed using DEA (Charnes et al., 1978). In this paper, we use an output-oriented Banker, Charnes and Cooper (BCC) model with variable return to scale (VRS) proposed by Banker et al. (1984). We use VRS since there is no reason to assume that the process of ST through the LUV instrument exhibits constant returns. We use an output-oriented model because we want to assess how participants can improve their ST skills performance by reaching higher levels of ST skill measures given the levels of inputs used through the LUV instrument.

4.1 | Input–output specifications

We use a functional system representation (Ropohl, 1999) to capture a participant's assessment of the short case study about the real-world shrinkage of Lake Urmia. Following the functional system point of view, inputs represent the level of effort (or a surrogate measure of the level of effort) to complete an exercise (through the LUV instrument) where ST skills are used, and output represents the ST outcomes achieved as part of the exercise. While each decision-making unit (DMU)[3] is viewed as a participant's LUV response, it is essential to note that a LUV response reflects a participant's level of ST skills at a specific point in time. It can be affected by various exogenous factors such as one's mood at that moment or one's understanding of the particular task. Nevertheless, in this study, we assume that all participants operate in relatively homogenous environments. That is, a participant is 'comparable' with another in terms of their vignette responses, their input–output specifications and the environments where the participants responded to the vignette.

According to the literature, there is no attempt to measure ST skills from a multi-dimensional perspective. Therefore, we propose two conceptual models that measure a participant's level of ST skills (Table 3.) The first DEA model uses the variables proposed in Davis et al. (2020). Total word count of the LUV response is chosen as the input variable, a resource (level of effort) that one uses in answering the vignette, along with the three ST output measures, that is, variables, causal links and loops. We refer to this model as the baseline model (i.e. BASE) to compare with the original LUV formulation that assesses performance.

Next, we propose our final model. The output variables are selected by adding two new output variables: middle nodes and connectivity to the previous output variables. We also drop variables due to its strong correlation with causal link (0.95)[4] and its direct consideration in all of our other output measures. The interpretation of the model is as follows: Given the total number of word count, the number of causal links, loops, middle nodes and connectivity, an instantiation of a LUV response is defined.

5 | RESULTS

5.1 | Influential observations: Step 1

By applying ROBPCA on three different variable combinations we obtain three different perspectives. First, when we only include input and output variables representing ST skill measures, the results yield an orthogonal cut-off distance of 1.64 and score cut-off distance of 2.72 (Figure 2). Six bad leverage points (red points) are found. These responses are identified as outliers because of their relatively large number of word count or Loops, which differentiates them from the rest of the sample. Second, when we considered all input, output and contextual variables (Figure 3), two LUV responses are found as bad leverage points. Lastly, when only contextual variables are considered (Figure 4), there are no bad leverage points found, indicating there are no outliers/extreme observations when considering only the contextual variables. Before discarding any observation, we completed a meta-analysis to make sure that an observation discarded is very different from the remainder of the dataset. In this case, we dropped the seven outliers obtained from Figures 2 and 3 and continue to the next step.

[3]In DEA, decision-making units (DMUs) are homogenous units that are being compared each one using multiple inputs and producing multiple outputs. Most often, DMUs may be companies, schools, hospitals, shops, bank branches, etc.

[4]The correlation matrix of the input and output are provided in Table A2 in Appendix A, Supporting Information.

TABLE 3 Variables specification for the LUV, BASE and multi-index models.[a]

Model	Input	Output				
	Word count	Variables	Links	Loops	Connectivity	Middle nodes
LUV	N/A	✓	✓	✓		
BASE (DEA)	✓	✓	✓	✓		
Multi-index (DEA)	✓		✓	✓	✓	✓

[a]We also test other conceptual specifications of output variables including the one representing detailed complexity (variables, links and middle node), dynamic complexity (loops, connectivity and middle Nodes). Strong correlation of 1 between the detailed complexity representation and the BASE model was found. Strong correlation of 0.81 between the dynamic complexity representation and the multi-index model was found. Hence, we only consider the BASE and multi-index model in our analysis.

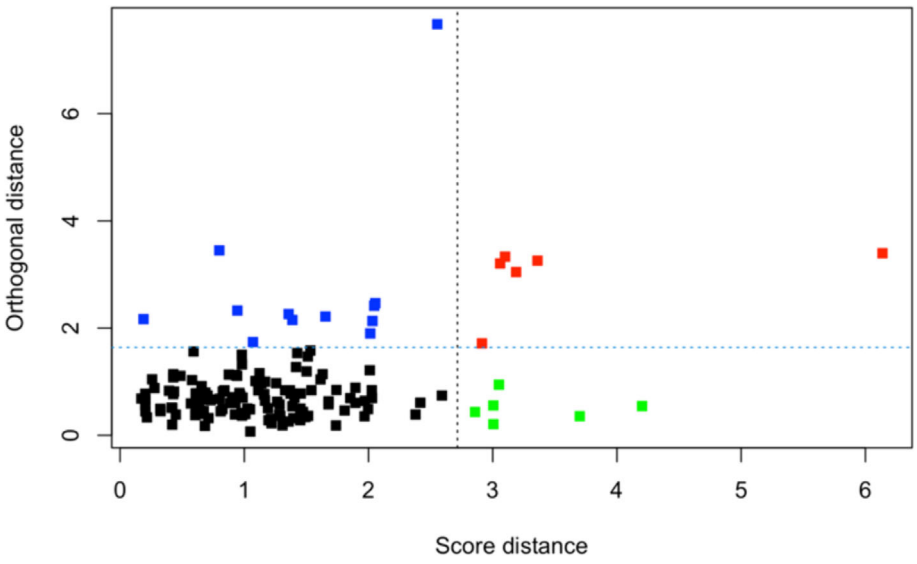

FIGURE 2 ROBPCA for outlier identifications on input and output variables. [Colour figure can be viewed at wileyonlinelibrary.com]

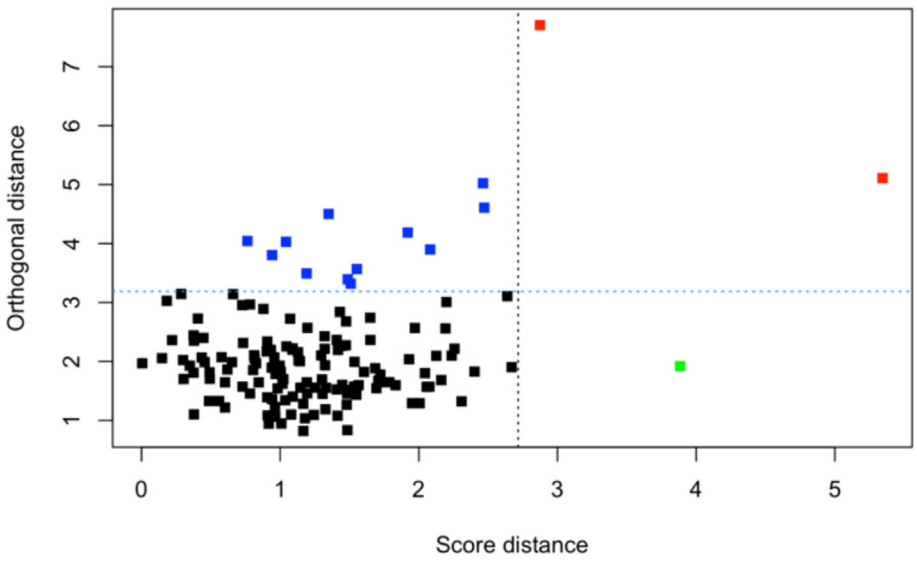

FIGURE 3 ROBPCA for outlier identifications on input, output and contextual variables. [Colour figure can be viewed at wileyonlinelibrary.com]

5.2 | The model results: Step 2

Step 2a starts by following the procedure proposed by Davis et al. (2020) to calculate the LUV score as a benchmark. We then compute the BASE and multi-index scores.[5] As opposed to the strong correlation (0.66) between LUV scores and word count, there is a weak

[5]Summary statistics for results of the three models is presented in Table A3 in Appendix A, Supporting Information.

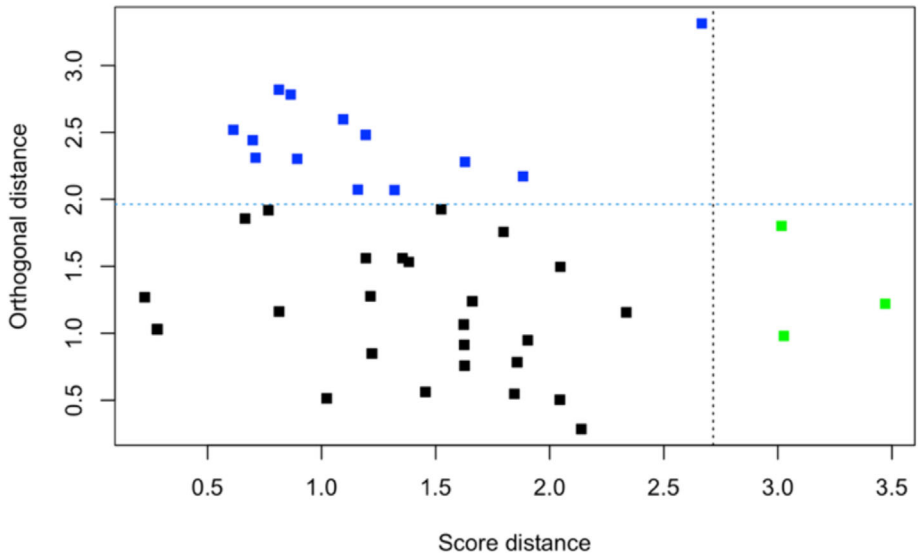

FIGURE 4 ROBPCA for outlier identifications on only contextual variables. [Colour figure can be viewed at wileyonlinelibrary.com]

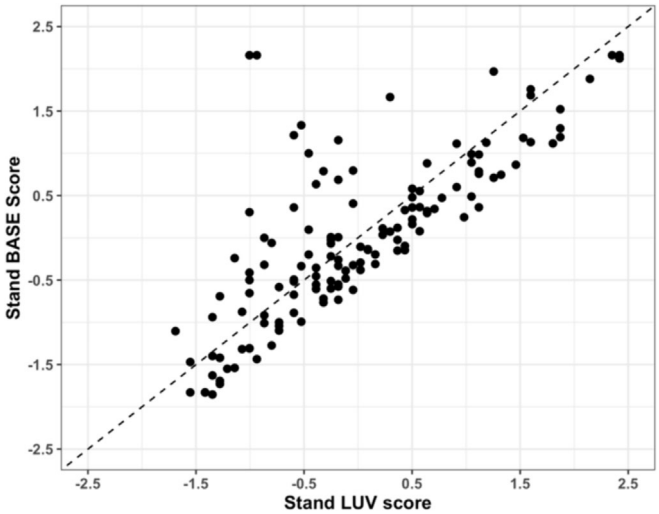

FIGURE 5 Relationship between LUV and BASE scores.

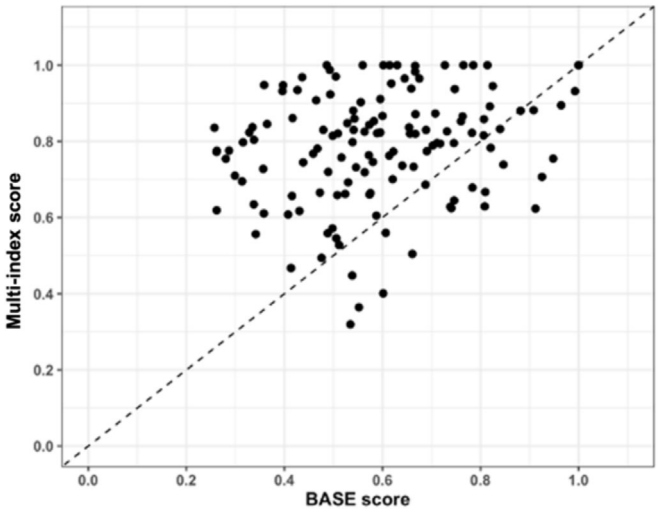

FIGURE 6 Relationship between BASE and multi-index scores.

correlation (0.23) between the BASE score and word count.[6] This indicates that DEA assigns a suitable weight for each response rather than solely relying on one's level of effort.

Next, a scatterplot compares the results of the three models (Figures 5 and 6). For better visualization, we normalized the score.[7] First, the BASE model identifies six LUV responses (i.e. vignette responses) as observations representing the highest ST skills, while the multi-index model identifies 17 vignette responses on the frontier. The multi-index model provides an average performance score of 0.79 that is higher than the average performance score of 0.61 of the BASE model. The LUV and BASE scores show a strong positive correlation of 0.79 (Figure 5). On the other hand, the scatter plot in Figure 6 suggests a weak correlation of 0.34 between BASE and multi-index score, due to the two additional variables, connectivity and middle nodes.

While most responses lie on the diagonal line, revealing the same score level in the overall distribution (Figure 5), few responses stand out from the others—the responses on the upper right corner show high scores in both models. However, the upper left corner reveals the distinction between the LUV linear combination-based method and the DEA-based method. These responses score high in the BASE model but low in the LUV model. While having relatively low word count compared to the others, these responses are associated with higher levels

[6]The correlation matrix is shown in Table A2 in the Supporting Information, Appendix A.

[7]We normalized the score by this function $z = (x - u)/\sigma$, where x is the original score, μ is the mean and σ is the standard deviation.

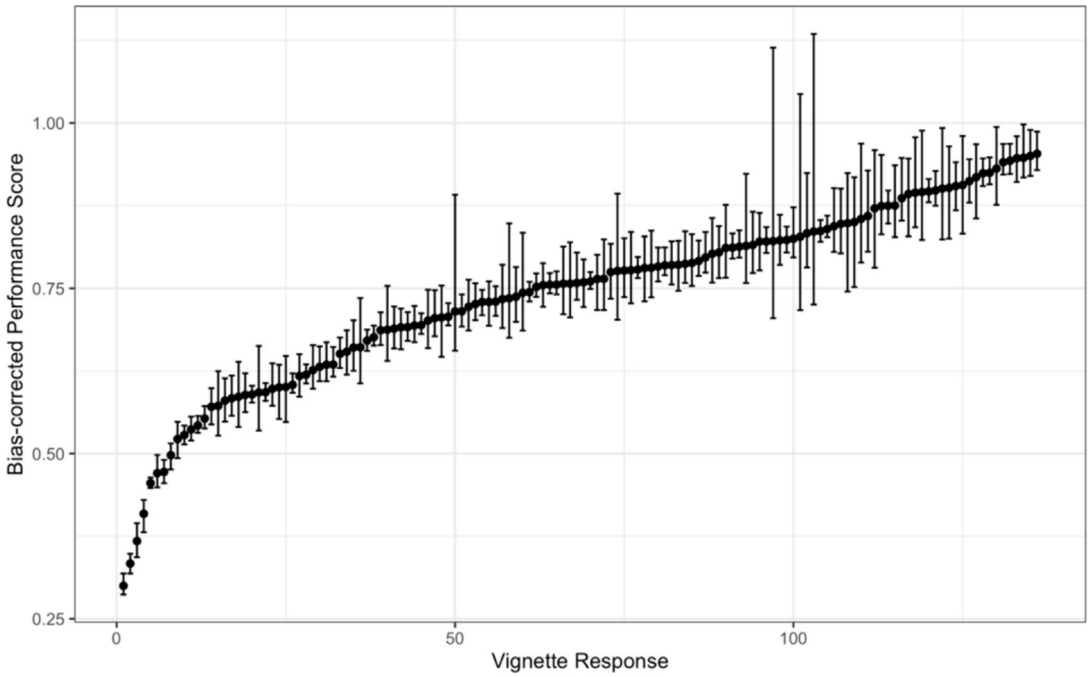

FIGURE 7 Bias-corrected performance score with 95% confidence intervals.

of ST skill measures when compared to their peers. These responses score high in the BASE model because the DEA approach does not penalize responses with lower Word Count.

In addition to the 'performance scores' obtained from DEA, a list of 'Peers' is provided for which each underperforming response should ideally emulate in order to improve its level of ST. This result may be used in a variety of ways. First, it provides low-performance responses with a list of responses they should attempt to emulate. Second, based on the frequency with which a high-performance response (i.e. performance score of 1) appears as a peer of interest, we can determine whether this response is suitable as a reference. A low frequency suggests that the response has extreme characteristics, which makes it an unsuitable peer (Athanassopoulos & Shale, 1997).

Around 25%–30% of high-performance vignette responses only appear as peers of interest only as little as five to nine times (Table A4 in Appendix A, Supporting Information). These responses are deemed to be high performance because of their 'extreme' characteristics, such as extreme (low/high) word count; therefore, should be treated with caution. In contrast, high-frequency responses may be treated as suitable observations to emulate.

In step 2a, we use an output-oriented Banker et al. (1984) model to obtain multi-index scores and estimate the ST skills' performance frontier. To address the bias arising from contextual variables, in step 2b, we use the two-staged bootstrapping approach proposed by (Simar & Wilson, 2007).

The result shows that the bias-corrected multi-index score is always lower than the naïve multi-index score since it accounts for the sampling noise originating from the contextual variables. Moreover, the plot of the bias-corrected multi-index scores and their confidence intervals suggests that higher performance responses have a larger bias (Figure 7), and we should treat these observations with care when they are used as peer observations.

5.3 | Contextual variables and their impact on the multi-index scores: Steps 3 and 4

In step 3, we ran regression models on the contextual variables to test if they significantly predict the LUV, multi-index and bias-corrected ulti-index scores. The regression results are shown in Table 4. It was found that only 'Nationality = international students' ($\beta = 8.54$, $p < 0.05$) significantly predicted the LUV scores. Second, when applying an OLS regression to predict the naïve multi-index scores, no contextual factors are found significant.

Third, we use the second-stage bootstrap approach to test the impact of the contextual variables on the bias-corrected multi-index scores (step 3b). We followed the algorithm proposed by (Simar & Wilson, 2007), a bootstrapped-truncated regression with 100 replications

TABLE 4 Regression results when considering the contextual variables as explanatory of performance scores.

	LUV score		Naïve multi-index score		Bias-corrected multi-index score	
Contextual variables	**Coeff.**	**Std. error**	**Coeff.**	**Std. error**	**Coeff.**	**Std. error**
(Intercept)	38.961	21.204	0.617	0.469	4.231	5.626
Gender	−0.555	1.380	0.021	0.031	−1.594	1.823
International	8.540	4.213*	0.083	0.093	−0.239	0.312
Age	−1.304	1.123	0.009	0.025	−0.422	0.413
Self-rate math	0.833	0.848	0.003	0.019	0.033	0.219
Feedback score	1.108	0.867	−0.021	0.019	0.359	0.261
N	135		135		135	
*R*2	0.073		0.02		N/A	
Adjusted *R*2	0.037		−0.02		N/A	

Note: Binary variables: female = 1, international = 1.
* < 0.05.

on both the first and second loops on the analysis at a = 0.05 and a = 0.10 significant levels. No significant factors are found impacting the bias-corrected multi-index scores. In this regression, the response variable was the inverse of the bias-corrected performance score. Therefore, a smaller independent variable indicates a better performance score. We did not find significant factors explaining the naïve multi-index scores and the bias-corrected multi-index scores. This may be due to the fact that the undergraduate student dataset has a higher homogeneity, since the study participants are all first-year engineering students without any training in ST.

5.4 | Replicability of the results

To test the generalizability for a different population group, we repeat the analysis for a sample of graduate students ($N = 30$). The results are reported in Appendix B. They are qualitatively consistent with the main results.

6 | DISCUSSION

In this paper, we propose a multi-dimensional index to evaluate ST skills from textual data. Our results show that our models are less affected by the length of responses (word count) compared to the LUV measure, indicating our measurement framework does not solely rely on participant's efforts. In addition, the multi-index scores take into account a more comprehensive ST skill set by adding the two additional measures, that is, middle nodes and connectivity. We would aspire to consider additional graph measures in future research. Furthermore, our results suggest that a few high-performance responses have larger biases, which should be treated with caution. Lastly, we discovered that there are no significant predictors of one's level of ST in the undergraduate dataset. Nevertheless, it should be noted that we also tested our proposed framework with an additional dataset of 30 graduate students. Detailed results are reported in Appendix B in the Supporting Information. In this dataset, we discovered that international students, self-reported math skills and older students demonstrate higher levels of ST skills. These results are in line with the finding of other studies in recent years (Davis et al., 2020). However, other contextual factors such as PhD student and training in systems are not found as significant predictors in the naïve and bias-corrected multi-index models. These inconsistencies need to be explored further by analysing additional datasets.

The innovation of our research lies in the determination of the appropriate input and output ST measures based on the discussed theoretical background and the generalizability of our method to any scenario where multiple ST skill measures need to be considered. The methodology employed allows us to engage in the determination of appropriate interventions that can arise from the significance (or not) of specific contextual factors and the identification of peers or best practices, along with the determination of performance targets. However, how our results could potentially translate into specific ST teaching guidelines is an open question and beyond the scope of this research paper. Nevertheless, it is a crucial topic since any performance measurement framework should offer suggestions as to where ST skills are lacking and how our curricula can be designed to address gaps.

While we have obtained some promising results, there is also a considerable amount of work that can be

done in the future. First, the dynamic changes of individuals' understanding of complexity need to be studied. A subset of the participants in this study are first-year engineering undergraduate students who have not decided their engineering major. Future work can include having the same set of students conduct the LUV task again during their senior year when they already have taken a number of engineering courses in their major. We can then examine the dynamic changes of the scores to see how their understanding of complexity changes over the years. Moreover, the study can be conducted with different vignettes that can describe various complex social problems. Examples include societal response to a pandemic, economic recession or social inequalities.

In addition, in this study, we use the most basic DEA model, the Banker et al. (1984) output-oriented model as a baseline for our model. Future assessments can be conducted using other DEA models that have different fundamental assumptions, such as the free disposal hull model (Deprins & Simar, 1984) where we relax the convexity assumption of DEA, additive (Ali & Seiford, 1993) and slack-based models (Tone, 2001) for which non-radial measures of performance are computed, or weight restriction models (Dyson et al., 2001) where we have the opportunity to weigh the various ST output skill measures based on expert opinion. Additionally, how these models can map to some of the basic notions of ST needs to be explored. Future work can include analysing additional sets of questions in the vignette to discover other significant contextual factors that impact one's level of ST skills. Last but not least, the process of the identification of variables and links is through a manual process, handed by coders. Future work may automate this process through the artificial intelligence, such as large language models, to reduce the time to hand-code text to mental maps (Wei et al., 2022).

It is important to acknowledge that our approach to, and definition of, ST was closer to the system dynamics school of thought. Assessment of individuals' ST skills as defined by other schools of thought is a future avenue of research. We invite researchers to examine other schools of thought, including but not limited to, critical ST (Jackson, 2016), and soft ST (Checkland & Haynes, 1994), or more perspectives as defined in general system theory (Von Bertalanffy, 1973).

7 | CONCLUSION

This study was a response to the challenge of measuring student's ST skills from their responses to a scenario documented in a text format. We proposed a multi-dimensional index that measures ST skills from such textual data. Our approach enables us to evaluate ST skills from various perspectives, including variables, causal links and feedback loops that a respondent considered when analysing a complex problem. While we have only tested our framework on the LUV, our index can be potentially applied to other complex case studies. Although our index is not a complete solution, this exploratory study marks a promising first step towards quantifying ST skills from textual data while considering its multi-dimensionality.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under NSF grant number 1824594. The statements, findings and conclusions are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CONFLICT OF INTEREST STATEMENT

The authors report there are no competing interests to declare.

ETHICS STATEMENT

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Institutional Review Board (IRB) of Virginia Polytechnic Institute and State University. IRB # 18-942.

ORCID

Ning-Yuan Georgia Liu https://orcid.org/0000-0002-1743-2040
Hesam Mahmoudi https://orcid.org/0000-0002-9524-6534
Konstantinos Triantis https://orcid.org/0000-0002-7407-290X
Navid Ghaffarzadegan https://orcid.org/0000-0003-3632-8588

REFERENCES

Ali, A. I., & Seiford, L. M. (1993). The mathematical programming approach to efficiency analysis. In H. O. Fried, C. A. Knox Lovell, & S. S. Schmidt (Eds.), *The measurement of productive efficiency: Techniques and applications* (pp. 120–159). Oxford Academic.

Andalib, M. A. (2018). Model-based analysis of diversity in higher education. PhD Dissertation, Virginia Tech, Department of Industrial and Systems Engineering.

Assaraf, O. B., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching: the Official Journal of the National Association for Research in Science Teaching*, *42*(5), 518–560. https://doi.org/10.1002/tea.20061

Athanassopoulos, A., & Shale, E. (1997). Assessing the comparative efficiency of higher education institutions in the UK by the

means of data envelopment analysis. *Education Economics*, *5*(2), 117–134. https://doi.org/10.1080/09645299700000011

Baghaei Lakeh, A., & Ghaffarzadegan, N. (2015). Does analytical thinking improve understanding of accumulation? *System Dynamics Review*, *31*(1–2), 46–65. https://doi.org/10.1002/sdr.1528

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*(9), 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078

Barlas, Y., & Diker, V. G. (2000). A dynamic simulation game (UNIGAME) for strategic university management. *Simulation & Gaming*, *31*(3), 331–358. https://doi.org/10.1177/104687810003100302

Bhatkoti, R., Triantis, K., Moglen, G. E., & Sabounchi, N. S. (2018). Performance assessment of a water supply system under the impact of climate change and droughts: Case study of the Washington metropolitan area. *Journal of Infrastructure Systems*, *24*(3). https://doi.org/10.1061/(asce)is.1943-555x.0000435

Black, L. J. (2013). When visuals are boundary objects in system dynamics work. *System Dynamics Review*, *29*(2), 70–86. https://doi.org/10.1002/sdr.1496

Cabrera, D., & Cabrera, L. (2019). What is systems thinking? In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_100-1

Cavaleri, S., & Sterman, J. D. (1997). Towards evaluation of systems-thinking interventions: A case study. *System Dynamics Review: the Journal of the System Dynamics Society*, *13*(2), 171–186. https://doi.org/10.1002/(SICI)1099-1727(199722)13:2<171::AID-SDR123>3.0.CO;2-9

Charnes, A., Cooper, W. W., Lewin, A. Y., & Seiford, L. M. (1994). Basic DEA models. In *Data envelopment analysis: Theory, methodology, and applications* (pp. 23–47). Springer. https://doi.org/10.1007/978-94-011-0637-5_2

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

Checkland, P. B., & Haynes, M. G. (1994). Varieties of systems thinking: The case of soft systems methodology. *System Dynamics Review*, *10*(2–3), 189–197. https://doi.org/10.1002/sdr.4260100207

Cooper, W. W., Kingyens, A. T., & Paradi, J. C. (2014). Two-stage financial risk tolerance assessment using data envelopment analysis. *European Journal of Operational Research*, *233*(1), 273–280. https://doi.org/10.1016/j.ejor.2013.08.030

Cooper, W. W., Seiford, L. M., & Zhu, J. (Eds.). (2011). *Handbook on data envelopment analysis* (Vol. 164). Springer. https://doi.org/10.1007/978-1-4419-6151-8

Cronin, S. J., Gaylord, D. R., Charley, D., Alloway, B. V., Wallez, S., & Esau, J. W. (2004). Participatory methods of incorporating scientific with traditional knowledge for volcanic hazard management on Ambae Island, Vanuatu. *Bulletin of Volcanology*, *66*, 652–668. https://doi.org/10.1007/s00445-004-0347-9

Darabi, N., Ebrahimvandi, A., Hosseinichimeh, N., & Triantis, K. (2021). A DEA evaluation of US states' healthcare systems in terms of their birth outcomes. *Expert Systems with Applications*, *182*, 115278. https://doi.org/10.1016/j.eswa.2021.115278

Deprins, D., & Simar, L. (1984). Measuring labor efficiency in post offices. In M. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The performance of public enterprises: Concepts and measurements*. North-Holland.

Davis, K., Ghaffarzadegan, N., Grohs, J., Grote, D., Hosseinichimeh, N., Knight, D., Mahmoudi, H., & Triantis, K. (2020). The Lake Urmia vignette: A tool to assess understanding of complexity in socio-environmental systems. *System Dynamics Review*, *36*(2), 191–222. https://doi.org/10.1002/sdr.1659

Davis, K. A., Grote, D., Mahmoudi, H., Perry, L., Ghaffarzadegan, N., Grohs, J., Hosseinichimeh, N.,Knight, D. B., & Triantis, K. (2023). Comparing self-report assessments and scenario-based assessments of systems thinking competence. *Journal of Science Education and Technology*, *32*(6), 793–813. https://doi.org/10.1007/s10956-023-10027-2

Dorani, K., Mortazavi, A., Dehdarian, A., Mahmoudi, H., Khandan, M., & Mashayekhi, A. N. (2015). Developing question sets to assess systems thinking skills. Proceedings of the 33rd International Conference of the System Dynamics Society, Cambridge, Massachusetts, USA.

Dugan, K. E., Mosyjowski, E. A., Daly, S. R., & Lattuca, L. R. (2022). Systems thinking assessments in engineering: A systematic literature review. *Systems Research and Behavioral Science*, *39*(4), 840–866. https://doi.org/10.1002/sres.2808

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, *132*(2), 245–259. https://doi.org/10.1016/S0377-2217(00)00149-1

Fordyce, D. (1988). The development of systems thinking in engineering education: An interdisciplinary model. *European Journal of Engineering Education*, *13*(3), 283–292. https://doi.org/10.1080/03043798808939427

Forrester, J. W. (1971). Counterintuitive behavior of social systems. *Theory and Decision*, *2*(2), 109–140. https://doi.org/10.1007/bf00148991

Forrester, J. W. (1997). Industrial dynamics. *Journal of the Operational Research Society*, *48*(10), 1037–1041. https://doi.org/10.1057/palgrave.jors.2600946

Ghaffarzadegan, N., & Larson, R. C. (2018). SD meets OR: A new synergy to address policy problems. *System Dynamics Review*, *34*(1–2), 327–353. https://doi.org/10.1002/sdr.1598

Gray, S., Sterling, E. J., Aminpour, P., Goralnik, L., Singer, A., Wei, C., Akabas, S., Jordan, R. C., Giabbanelli, P. J., Hodbod, J., Betley, E., & Norris, P. (2019). Assessing (social-ecological) systems thinking by evaluating cognitive maps. *Sustainability*, *11*(20), 20. https://doi.org/10.3390/su11205753

Grohs, J. R., Kirk, G. R., Soledad, M. M., & Knight, D. B. (2018). Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems. *Thinking Skills and Creativity*, *28*, 110–130. https://doi.org/10.1016/j.tsc.2018.03.003

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*, 769–794. https://doi.org/10.1037/pspi0000155

Haque, S., Mahmoudi, H., Ghaffarzadegan, N., & Triantis, K. (2023). Mental models, cognitive maps, and the challenge of

quantitative analysis of their network representations. *System Dynamics Review*, *39*, 152–170. https://doi.org/10.1002/sdr.1729

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford University Press. https://doi.org/10.1093/oso/9780195130133.001.0001

Hendijani, R. (2021). Analytical thinking, Little's law understanding, and stock-flow performance: Two empirical studies. *System Dynamics Review*, *37*(2–3), 99–125. https://doi.org/10.1002/sdr.1685

Herrera-Restrepo, O., Triantis, K., Trainor, J., Murray-Tuite,P., & Edara, P. (2016). A multi-perspective dynamic network performance efficiency measurement of an evacuation: A dynamic network-DEA approach. *Omega*, *60*, 45–59. https://doi.org/10.1016/j.omega.2015.04.019

Howie, E., Sy, S., Ford, L., & Vicente, K. J. (2000). Human–computer interface design can reduce misperceptions of feedback. *System Dynamics Review: the Journal of the System Dynamics Society*, *16*(3), 151–171. https://doi.org/10.1002/1099-1727(200023)16:3<151::AID-SDR191>3.0.CO;2-0

Hubert, M., & Engelen, S. (2004). Robust PCA and classification in biosciences. *Bioinformatics*, *20*(11), 1728–1736. https://doi.org/10.1093/bioinformatics/bth158

Huz, S., Andersen, D. F., Richardson, G. P., & Boothroyd, R. (1997). A framework for evaluating systems thinking interventions: An experimental approach to mental health system change. *System Dynamics Review: the Journal of the System Dynamics Society*, *13*(2), 149–169. https://doi.org/10.1002/(SICI)1099-1727(199722)13:2<149::AID-SDR122>3.0.CO;2-S

Jackson, M. C. (2016). *Systems thinking: Creative holism for managers*. John Wiley & Sons, Inc.

Kim, H. (2009). In search of a mental model-like concept for group-level modeling. *System Dynamics Review*, *25*(3), 207–223. https://doi.org/10.1002/sdr.422

Kim, H., & Andersen, D. F. (2012). Building confidence in causal maps generated from purposive text data: Mapping transcripts of the federal reserve. *System Dynamics Review*, *28*(4), 311–328. https://doi.org/10.1002/sdr.1480

Kunc, M. H., & Morecroft, J. D. W. (2007). Competitive dynamics and gaming simulation: Lessons from a fishing industry simulator. *Journal of the Operational Research Society*, *58*(9), 1146–1155. https://doi.org/10.1057/palgrave.jors.2602246

Lane, D. C. (1995). On a resurgence of management simulations and games. *Journal of the Operational Research Society*, *46*(5), 604–625. https://doi.org/10.1057/jors.1995.86

Levy, M. A., Lubell, M. N., & McRoberts, N. (2018). The structure of mental models of sustainable agriculture. *Nature Sustainability*, *1*(8), 413–420. https://doi.org/10.1038/s41893-018-0116-y

Maani, K. E., & Maharaj, V. (2004). Links between systems thinking and complex decision making. *System Dynamics Review: the Journal of the System Dynamics Society*, *20*(1), 21–48. https://doi.org/10.1002/sdr.281

Mahmoudi, H., Dorani, K., Dehdarian, A., Khandan, M., & Mashayekhi, A. N. (2019). Does systems thinking assessment demand a revised definition of systems thinking? 37th International Conference of the System Dynamics Society, Albuquerque, NM, USA.

Montibeller, G., & Belton, V. (2006). Causal maps and the evaluation of decision options—A review. *Journal of the Operational Research Society*, *57*(7), 779–791. https://doi.org/10.1057/palgrave.jors.2602214

Moxnes, E. (2000). Not only the tragedy of the commons: Misperceptions of feedback and policies for sustainable development. *System Dynamics Review: the Journal of the System Dynamics Society*, *16*(4), 325–348. https://doi.org/10.1002/sdr.201

Moxnes, E. (2004). Misperceptions of basic dynamics: The case of renewable resource management. *System Dynamics Review*, *20*(2), 139–162. https://doi.org/10.1002/sdr.289

Naugle, A., Verzi, S., Lakkaraju, K., Swiler, L., Warrender, C., Bernard, M., & Romero, V. (2021). Feedback density and causal complexity of simulation model structure. *Journal of Simulation*, *0*(0), 1–11. https://doi.org/10.1080/17477778.2021.1982653

Nehdi, M., & Rehan, R. (2007). Raising the bar for civil engineering education: Systems thinking approach. *Journal of Professional Issues in Engineering Education and Practice*, *133*(2), 116–125. https://doi.org/10.1061/(ASCE)1052-3928(2007)133:2(116)

Özgün, O., & Barlas, Y. (2015). Effects of systemic complexity factors on task difficulty in a stock management game. *System Dynamics Review*, *31*(3), 115–146. https://doi.org/10.1002/sdr.1543

Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. *New Directions for Institutional Research*, *2011*(150), 41–58. https://doi.org/10.1002/ir.388

Plate, R. (2010). Assessing individuals' understanding of nonlinear causal structures in complex systems. *System Dynamics Review*, *26*(1), 19–33. https://doi.org/10.1002/sdr.432

Porter, S. R. (2011). Do college student surveys have any validity? *Review of Higher Education: Journal of the Association for the Study of Higher Education*, *35*, 45–76. https://doi.org/10.1353/rhe.2011.0034

Richardson, G. P. (1994). Systems thinkers, systems thinking.*System Dynamics Review*, *10*(2–3), 95–99. https://doi.org/10.1002/sdr.4260100202

Richardson, G. P. (2011). Reflections on the foundations of system dynamics. *System Dynamics Review*, *27*(3), 219–243. https://doi.org/10.1002/sdr.462

Richmond, B. (1993). Systems thinking: Critical thinking skills for the 1990s and beyond. *System Dynamics Review*, *9*(2), 113–133. https://doi.org/10.1002/sdr.4260090203

Richmond, B. (2000). *The "thinking" in systems thinking: Seven essential skills*. Pegasus Communications.

Ropohl, G. (1999). Philosophy of socio-technical systems. *Society for Philosophy and Technology Quarterly Electronic Journal*, *4*(3), 186–194. https://doi.org/10.5840/techne19994311

Samoylova, E. (2014). Virtual world of computer games: Reality or illusion? *Procedia-Social and Behavioral Sciences*, *149*, 842–845. https://doi.org/10.1016/j.sbspro.2014.08.324

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, *136*(1), 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009

Stave, K., & Hopper, M. (2007). What constitutes systems thinking? A proposed taxonomy. 25th International Conference of the System Dynamics Society.

Sterman, J. D. (1989a). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes*, *43*(3), 301–335. https://doi.org/10.1016/0749-5978(89)90041-1

Sterman, J. D. (1989b). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, *35*(3), 321–339. https://doi.org/10.1287/mnsc.35.3.321

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Irwin/McGraw-Hill.

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

Triantis, K., Sarayia, D., & Seaver, B. (2010). Using multivariate methods to incorporate environmental variables for local and global efficiency performance analysis. *INFOR: Information Systems and Operational Research*, *48*(1), 39–52. https://doi.org/10.3138/infor.48.1.039

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, *130*(3), 498–509. https://doi.org/10.1016/S0377-2217(99)00407-5

Von Bertalanffy, L. (1973). The meaning of general system theory. *General System Theory: Foundations, Development, Applications*, *30*, 53.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. arXiv. arXiv:2206.07682. https://doi.org/10.48550/arXiv.2206.07682

Wolstenholme, E. F. (1993). A case study in community care using systems thinking. *Journal of the Operational Research Society*, *44*(9), 925–934. https://doi.org/10.1057/jors.1993.160

Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, R. J. (2005). Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*, *42*(2), 166–184. https://doi.org/10.1002/tea.20049

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu, N.-Y. G., Mahmoudi, H., Triantis, K., & Ghaffarzadegan, N. (2024). A multi-dimensional index of evaluating systems thinking skills from textual data. *Systems Research and Behavioral Science*, 1–15. https://doi.org/10.1002/sres.3033