Measuring Systems Thinking with a Multidimensional Inventory of Systems Thinking Skills

Joe A. Wasserman

Author Note

Correspondence concerning this paper should be addressed to Joe A. Wasserman. E-mail: joe.wasserman@gmail.com

Measuring Systems Thinking with a Multidimensional Inventory of Systems Thinking Skills

Systems thinking—the ability to understand, make predictions about, and intervene in complex systems—has been identified as a crucial skill for learners and practitioners across many disciplines, from sciences and engineering (National Research Council, 2012) to management and organizational studies (Checkland, 1999) to social psychology (Levine & Doyle, 2002) and social sciences (Garson, 2009). Because systems are diverse, they must be conceptualized broadly to match: systems are groups of things that are interrelated (Bertalanffy, 2015), including but not limited to people in social systems, firms in economic systems, nation states in political systems, various species in ecological systems, or organelles in cellular systems. Given the importance of systems thinking to so many disciplines, there is a need for measures of systems thinking that are not tied to any particular domain. To date, measurement attempts have been relatively focused, reflecting either a narrow range of systems thinking competencies or expertise in a particular domain. After characterizing systems thinking in greater detail and reviewing existing measures of systems thinking, this study integrates existing, easy-to-score measures into a more complete, domain-independent, multi-dimensional measure of systems thinking skills.

Review of Literature

Given the breadth of scholarship that has laid claim to systems thinking, the term has been derided as having come "to mean little more than thinking about systems, talking about systems, and acknowledging that systems are important" (Forrester, 1994, p. 252). Nevertheless, more specific understandings of systems thinking have since been formalized as typologies of systems thinking. These typologies, reviewed below, are useful for researchers and practitioners in that they (a) provide structure to the otherwise potentially nebulous concept of systems thinking and (b) identify a range of more discrete abilities that can all be considered part of systems thinking more broadly.

Systems Thinking: Definitions and Typologies

The diverse literature on systems thinking contains multiple definitions, conceptualizations, and descriptions of systems thinking (e.g., Meadows, 2008; Senge, 2006). Given this conceptual diversity, it is productive to synthesize the literature to identify commonalities among systems thinking competencies. Toward this end, we focused on two typologies. The first, Assaraf and Orion's (2005) list of eight discrete systems thinking skills, was selected for its established influence, having been cited 380 times as of the time of writing ("Google scholar," n.d.). The second, Stave and Hopper's (2007) taxonomy of systems thinking, was selected for having drawn on a framework that is well-established in the learning sciences: Anderson and Krathwohl's (2001) revision of Bloom's taxonomy of educational objectives.

Assaraf and Orion (2005) identified eight systems thinking skills within a three-tiered hierarchy. The lowest level of this hierarchy focuses on identifying system components (skill 1), the second on synthesizing these components (skills 2–5), and the third on implementing systems thinking skills (skills 6–8). This skills are as follows.

1. "The ability to identify the components of a system and processes within the system" (p. 523) comprises two distinct skills, (a) identifying the entities that are crucial to a given system and (b) identifying the processes through which these entities change over time. For example, an ecological system might include a species' population (component) and their increase via reproduction (process).
2. "The ability to identify relationships among the system's components" (p. 523) is distinct from the first skill in that it emphasizes *interrelations* among entities, as opposed to

processes that occur within a given entity. Continuing the ecological system example, a relationship might include predation of one species by another.

3. “The ability to identify dynamic relationships within the system” (p. 523) entails recognizing not only that entities are interrelated, but that these interrelations cause changes to entities over time.
4. “The ability to organize the systems’ components and processes within a framework of relationships” (p. 523) integrates the first three systems thinking skills. Whereas the first three skills emphasize identification, the third focuses on the organization or integration of entities, processes, and interrelations into a holistic framework or system. Within an ecological system, such a framework might look like a food chain that includes all of the resources on which each species depends. These first four systems thinking skills are closely tied to the aforementioned conceptualization of systems as groups of interconnected entities (see Bertalanffy, 2015).
5. “The ability to understand the cyclic nature of systems” (p. 523) means recognizing the feedback loops within systems through which entities mutually influence each other over time, rather than proceeding from a starting point to an ending point. Balancing dynamics of predator-prey relationships are archetypical of these cyclic feedback loops: as predator populations grow, prey populations shrink due to predation, but as prey populations shrink, predator populations shrink in turn due to scarcity of food.
6. “The ability to make generalizations” (p. 523) involves synthesizing understandings of entities, processes, or interrelations within a given system, or of whole systems themselves. These generalizations might include classifying entities or interrelations within a system, for example by realizing the similarities of predation and reproduction at

each point in a food chain. Generalizing about a system more holistically might involve characterizing that system in terms of dynamic changes over time (e.g., fluctuations of population sizes), feedback loops among interrelated entities (e.g., the mutual dependence of predator-prey populations on each other), or temporal delays embedded in these interrelationships and processes (e.g., gestation periods that cause predators' population to grow after prey populations have already declined due to predation).

7. "Understanding the hidden dimensions of the system" (p. 523) involves recognizing that the entities, interrelations, and processes within a system may not be immediately apparent from surface inspection. In an ecological system, for example, decomposition by fungi and bacteria transform plant and animal matter into nutrient resources that are consumable by plants, but decomposition occurs at physical and temporal scales not readily observable by humans.
8. "Thinking temporally: retrospection and prediction" (p. 523) entails understanding that a given state of a system is a product of past interactions and processes, while future system states will be due to present dynamics.

Stave and Hopper (2007) developed a seven-tiered, hierarchical taxonomy of systems thinking that they aligned with Anderson and Krathwohl's (2001) revision of Bloom's taxonomy of cognitive processes of remembering, understanding, applying, analyzing, evaluating, and creating. Although their taxonomy overlaps in part with Assaraf and Orion (2005), it is novel in its alignment with an existing, well-established taxonomy of learning and describes additional systems thinking skills. This taxonomy of systems thinking skills (adapted from K. Stave & Hopper, 2007, Figure 4)—and their corresponding cognitive processes from Anderson and Krathwohl (2001)—are as follows.

1. *Recognizing interconnections* (remembering) includes "seeing the whole system, understanding how parts relate to and make up wholes, and recognizing emergent properties" (p. 14) and maps onto Assaraf and Orion's (2005) skills one, two, and three, and is closely related to the conceptualization of systems as groups of interrelated entities (Bertalanffy, 2015). It more specifically involves recognizing the manner in which a system is comprised of its component parts, as well as the way in which those parts are interrelated to form the whole system.
2. *Identifying feedback* (remembering) includes "recognizing and identifying interconnections and feedback" (p. 14) and maps onto Assaraf and Orion's (2005) skills two, three, and five. This skill entails recognizing the causal relationships among entities in a system, including indirect chains of causal relationships and feedback loops.
3. *Understanding dynamic behavior* (understanding) includes "understanding the relationship between feedback and behavior, including delays" (p. 14) and maps onto Assaraf and Orion's (2005) skills three, five, and six. Beyond recognizing interrelationships, this systems thinking skill involves understanding how system structures—such as feedback loops and delays—produce behaviors that are characteristic of that system.
4. *Differentiating types of variables and flows* (understanding) includes "understanding the difference between rates and levels" (p. 14) and is not contained in Assaraf and Orion (2005). This skill entails quantitatively characterizing entities and interrelationships of a system in terms of stocks and flows. Stocks are quantities of some material or information that have accumulated in a particular place, whereas flows are movements of this material or information into, out of, or between stocks (Meadows, 2008). Stocks are

characterized quantitatively in terms of their levels, or how much material or information they currently hold. Flows are characterized by rates, or the speed at which material or information moves through them. A bathtub provides a basic example of a stock (the water held in the tub) and two flows: a faucet that controls the water's rate of in-flow of into the tub and a drain that controls the water's rate of out-flow from the tub.

5. *Using conceptual models* (applying/analyzing) includes "using general systems principles to explain an observation" (p. 14) and maps onto Assaraf and Orion's (2005) skills six and eight. This skill involves applying general systems concepts to observations of particular systems in order to explain how system structures produced a particular system state.
6. *Creating simulation models* (applying/analyzing/evaluating/creating) includes "describing connections in mathematical terms and using both qualitative and quantitative variables" (p. 14) and is not contained in Assaraf and Orion (2005). This advanced systems thinking skill incorporates all of the prior skills and entails creating simulations of a particular system by formalizing a description of that system's entities and their interrelations. Creating simulations also involves comparing them to an external standard, such as observations of a system, other formal models, or other simulations.
7. *Testing policies* (applying/analyzing/evaluating/creating) includes "using simulation to test hypotheses and develop policies" (p. 14) and is not contained in Assaraf and Orion (2005). This advanced systems thinking skill may or may not involve creating simulations, but otherwise incorporates all of the prior skills. Testing policies entails using simulations to identify potential changes to a system that would produce desirable future states of the system. For example, a forestry service overrun with deer that are

consuming too much flora may use simulations to test a number of solutions to this problem, such as releasing wolves into the area or allowing hunting of deer during certain seasons.

Systems Thinking Measurement

A number of measurement instruments of systems thinking that reflect various aspects of the aforementioned systems thinking skills have been developed. These measures are diverse, including both closed- and open-ended items; graphical, numeric, and textual responses; and a range of content area specificity from domain-dependent to domain-independent. To maximize the utility of systems thinking measures, we propose that instruments should meet the following seven criteria: (a) reflect a wide range of systems thinking skills, (b) capture a wide range of systems thinking abilities, (c) not require knowledge of a particular content area—i.e., be domain-independent, (d) be easy to administer, (e) be easy to score, and (f) be as short possible so as to minimize burden on participants while satisfying the preceding criteria. The following review of systems thinking measures focuses on measures that meet most or all of these criteria.

Stocks and flows. An early attempt at measuring domain-independent systems thinking ability—and understanding of stocks and flows in particular—involved three open-ended, domain-independent graphing tasks (Booth Sweeney & Sterman, 2000). Each item graphically and textually depicted a stock and flow of some material over time, with two items showing different patterns of inflow and outflow, and the third incorporating a temporal delay and one feedback loop. Responses were elicited as freehand-drawn graphs and scored according to seven-to-eight characteristics. Among primarily MBA MIT students, performance on these items was rather poor, ranging from an average of 41% correct to 77% correct (Booth Sweeney & Sterman, 2000). Among high schoolers, undergraduates, and masters students, average response correctness for the two items without a delay or feedback loop has typically ranged from about 40% to about 80% (for a review, see Kapmeier, Happach, & Tilebein, 2017).

The aforementioned graphing-response measures of stock-and-flow understanding have been adapted as closed-ended, domain-independent, numeric-response items (Ossimitz, 2002; Sterman, 2002). In these closed-ended versions of stock-and-flow tasks, items graphically depict one inflow and one outflow, eliciting responses by asking participants to identify the moments of greatest inflow, greatest outflow, greatest stock, and least stock. The first two of these questions measure graph interpretation, while the latter two are more closely aligned with understand stocks and flows. Across multiple student samples, from high school to undergraduate to graduate, the proportion of correct responses to the latter two questions range from roughly 25% to 40% (Kasperidus, Langfelder, & Biber, 2006; Ossimitz, 2002; Sterman, 2002).

Both open-ended and closed-ended items measuring stock-and-flow understanding have been adapted, validated, and applied in a range of research settings. Poor performance on these items appears not to be attributable to task complexity, mode of numeric data presentation, the cover story introducing the task, or motivation to provide correct responses, but instead to the correlation heuristic, which involves treating stocks as though they were directly proportional to inflows (Cronin, Gonzalez, & Sterman, 2009; see also Kapmeier et al., 2017). Some item format characteristics have been found to improve performance on these stock-and-flow measures, however. Specifically, providing information about initial stocks and requesting responses as general qualitative statements as opposed to specific numeric values (Fischer & Degen, 2012), as about general trends with a tabular data presentation as opposed to otherwise (Fischer, Kapmeier, Tabacaru, & Kopainsky, 2015), and forewarning individuals about the difficulty of stock-and-flow questions (Baghaei Lakeh & Ghaffarzadegan, 2016) have all been found to improve individuals' performances.

Beyond item format, certain individual differences appear to be associated with performance on stock-and-flow measures. Specifically, an analytical thinking style—but not a global processing style—have been found to be associated with greater performance on closed-ended stock-and-flow tasks (Weinhardt, Hendijani, Harman, Steel, & Gonzalez, 2015). Additionally, engineering majors perform better than business majors (Kapmeier et al., 2017) and forestry majors better than sustainable resource management majors (Kasperidus et al., 2006). Within groups of business and engineering majors, however, differences in performance have not been found between to significantly differ (Kapmeier et al., 2017).

Despite aforementioned poor performances on measures of stock-and-flow understanding, this ability can be improved through training. Both a crash course on stocks and flows (Kainz & Ossimitz, 2002) and introductory system dynamics courses (Pala & Vennix, 2005; Sterman, 2010) have been found to improve individuals' responses to these items. Smaller interventions, such as interacting with a simulation of environmental accumulation (K. A. Stave, Beck, & Galvan, 2015) or a game about vaccinating populations (Kaufman & Flanagan, 2016) have also led to improvements in understandings of stocks and flows.

Mixed methods measures of systems thinking skills. A number of more domain-dependent, closed- and open-ended measures of systems thinking skills have been developed in the context of primary science education in earth sciences (Ben-Zvi Assaraf & Orion, 2005) and biology (Ben-Zvi Assaraf, Dodick, & Tripto, 2013). These measures include closed-ended true-or-false questions about system components, as well as simple, dynamic, and cyclical relationships; open-ended system drawings and concept maps that integrate system components and relationships into more-or-less coherent frameworks; open-ended word association tasks;

semi-structured interviews on making generalizations about systems, identifying hidden aspects of systems, and generating explanations and predictions; and a semi-structured repertory grid in which participants generate constructs with which they subsequently rate (Ben-Zvi Assaraf & Orion, 2005). Analyses of these diverse measures have provided convergent evidence that systems thinking can be cultivated among fourth grade students (Ben-Zvi Assaraf & Orion, 2010b) and junior high school students (Ben-Zvi Assaraf & Orion, 2005). Furthermore, these improvements can persist—to varying degrees—years later (Ben-Zvi Assaraf & Orion, 2010a). The open-ended and semi-structured word association, repertory grid, and concept mapping measures have been applied to high schoolers' understandings of human bodies as systems, finding that after completing their human body curriculum, high school students primarily remained at the most basic, component-level of systems thinking (Ben-Zvi Assaraf et al., 2013). The closed-ended, true-false measures have been applied to teachers' understandings of the water cycle, finding that although in-service teachers have greater low-level systems thinking abilities than pre-service teachers, overall, teachers found systems thinking difficult (Lee, Jones, & Chesnutt, 2017).

Systems Thinking Assessment. The Systems Thinking Assessment (STA) is a closed-ended, domain-independent multiple choice test that was iteratively and cyclically developed to measure middle school students' systems thinking abilities (K. C. Constantinide, 2015). This 52-item test was developed in Greek and measures identification of system components, reasoning about causal relationships, recognizing how system structures produce system behaviors, and reasoning about flows of matter that include feedback loops (K. Constantinide, Michaelides, & Constantinou, 2014).

Complex Systems Concepts Inventory. The Complex Systems Concepts Inventory

(CSCI) is a domain-independent measure of understanding general complex systems concepts that includes both closed-ended multiple choice and open-ended free response items (Tullis & Goldstone, 2017). Complex systems are characterized by individual actors or agents—whether animals, cells, or humans—from whose relatively simple interactions emerge complex, macroscopic patterns (Mitchell, 2009). As such, the CSCI measures understandings of general systems properties such as emergence, feedback, and self-organization; explaining system behaviors; and making predictions about complex systems. Across two groups of high school students, instruction in creating system simulations with agent-based modeling has been found to improve students' understandings of complex systems concepts (Tullis & Goldstone, 2017).

Other measures. Other systems thinking measurement instruments exist but do not meet the aforementioned seven criteria for systems thinking measure utility. For example, semi-structured interviews (Booth Sweeney & Sterman, 2007) are time-consuming both to administer and to score. Concept maps (Brandstädter, Harms, & Großschedl, 2012) and other open-ended textual and diagrammatic systems thinking measures (Evagorou, Korfiatis, Nicolaou, & Constantinou, 2009; Riess & Mischo, 2010) are domain-dependent, difficult to score, and are time-consuming for participants to create. For the aforementioned reasons, these measurement instruments were not included in this study.

Systems Thinking Skills and Systems Thinking Measures

It is possible to categorize the various aforementioned systems thinking measures according to the particular systems thinking skills they measure (e.g., Ben-Zvi Assaraf & Orion, 2005; Hopper & Stave, 2008). Based on the aforementioned review of systems thinking skills typologies, the reviewed measures of systems thinking that are closed-ended or relatively easy to score were categorized (see Table 1). As can be seen, each existing measure reflects a subset of systems thinking skills. In combination, however, these measures reflect the majority of basic to moderate-level systems thinking competencies. Therefore, by combining these measures, it should be possible to arrive at a multi-dimensional measure of systems thinking skills that captures this range of competencies. The two categorizations of measures in Table 1 present competing hypothesized factor structures for the systems thinking measures examined in this study.

Table 1

Measures of systems thinking categorized by the skills they are hypothesized to measure

	Stave and Hopper (2007)						
Assaraf and Orion (2005)	1. Recognize interconnections	2. Identify feedback	3. Understand dynamic behavior	4. Differentiate types of variables and flows	5. Use conceptual models	6. Create simulation models	7. Test policies
1. Identify components and processes	STA	STA	STA	STA			
2. Identify relationships	GDN, STA	STA	STA	STA			
3. Identify dynamic relationships	GDN, STA	STA	STA	STA			
4. Organize within framework							
5. Understand cycles	STA	CTQ, STA	STA	CTQ, STA			
6. Make generalizations			CSCI				
7. Understand hidden dimensions			CSCI				
8. Retrospection and prediction			CSCI				
N/A				DS			

Note. GDN: Groundwater system dynamic nature questionnaire (Ben-Zvi Assaraf & Orion, 2005); CTQ: Cyclic thinking questionnaire (Ben-Zvi Assaraf & Orion, 2005); CSCI: Complex Systems Concepts Inventory (Tullis & Goldstone, 2017); DS: Department store task (Sterman, 2002); STA: Systems Thinking Assessment (Constantinide, 2015).

Method

Participants

Participants were 465 English-speaking individuals age 18 and over. Of the 452 who reported their age, ages ranged from 18 to 97 years (*M* = 27.1, *SD* = 12.6). Of the 458 who reported their sex, 189 (41.3%) were male, 251 (54.8%) were female, and 12 (2.6%) were trans or nonbinary. Of the 454 who reported their race, 374 (82.4%) were White, and the next most common race was Asian (*n* = 26, 5.7%). Of the 456 who reported their highest education, 244 (53.5%) had completed high school or some college, 130 (28.5%) had completed a two- or four-year degree, and 81 (17.8%) had obtained a graduate or professional degree.

Procedure

Following approval from the university's Institutional Review Board, participants were recruited from a student email list at a large Mid-Atlantic university, social media, online forums, and online databases of psychology studies so as to obtain participants with a wide range of systems thinking abilities. Participants completed an online questionnaire containing items measuring endorsement of systems concepts, systems thinking skills, and demographics. Within the first two groups of measures, presentation order was randomized and time to complete each page of the survey was timed.

Measures

Systems thinking skills. A range of systems thinking skills were measured using four existing measures of systems thinking that included both open- and closed-ended items, to be combined into a single Multidimensional Inventory of Systems Thinking Skills (MISTS).

Department store task. Systems thinking skills were measured with the department store task (DS; Sterman, 2002), which contains four numeric free-response questions on a graph of flows of individuals entering and exiting a department store over a 30-minute period. Sample items included, “During which minute did the most people leave the store?,” and “During which minute were the most people in the store?” DS has been described and discussed in the preceding literature review. Answers within ±1 of the correct response were scored as correct.

Groundwater System Dynamic Nature Questionnaire. Systems thinking skills were measured with the Groundwater System Dynamic Nature Questionnaire (GDN; Ben-Zvi Assaraf & Orion, 2005), which contains eight true-or-false (*agree*, *uncertain*, or *disagree*) items on the interrelationships between water and other elements of the hydro-system. Sample items included, “Rocks don’t influence the composition of the water that penetrates them,” and “Many factories have their sewage flow into streams, thus polluting the water we drink.” GDN has been described and discussed in the preceding literature review.

Cyclic Thinking Questionnaire. Systems thinking skills were measured with the Cyclic Thinking Questionnaire (CTQ; Ben-Zvi Assaraf & Orion, 2005), which contains six true-or-false (*agree*, *uncertain*, or *disagree*) items on the cyclical nature of the stocks and flows of the hydro-system. Sample items included, “Clouds are the starting point of the water cycle and the tap at home is its end point,” and “The amount of water that evaporates into the atmosphere from the entire surface of the earth is not equal to the amount of rain that falls on the earth’s surface.” CTQ has been described and discussed in the preceding literature review.

Systems Thinking Assessment. Systems thinking skills were measured with the Systems Thinking Assessment (STA; K. C. Constantinide, 2015; K. Constantinide et al., 2014) translated from Greek into English by the creator and slightly revised by the first author for clarity and cultural context, which contains 29, 4-option multiple choice questions on system components, interrelationships among components, temporal causal dynamics, and feedback loops in everyday contexts. Because it was developed for 10–14 year olds, STA was written in plain language. Sample items included, “Which parts are ABSOLUTELY NECESSARY for a bicycle to roll when someone pushes it?,” “A brand of coffee in a supermarket becomes more and more expensive each passing week. What are most of the customers likely to do in a few weeks?,” and “Could the water that leaves a bathtub come back to the same house as drinking water?” STA has been described and discussed in the preceding literature review.

Complex Systems Concepts Inventory. Systems thinking skills were measured with the Complex Systems Concepts Inventory (CSCI; Tullis & Goldstone, 2017), which is comprised of two versions of analogous questions. Each version contains one closed-ended puzzle that asks participants to implement system rules; one closed-ended, 4-option multiple-choice question on predicting the outcome of system dynamics; and four open-ended questions on explaining system dynamics. Sample open-ended questions include “A pattern of ridges and troughs can be formed when varnish begins to wrinkle and lift off of wood, as shown below. How can this complex pattern occur?,” and “Some groups of fireflies will begin to synchronize their flashing after spending some time together in an area. How might large groups synchronize their flashing?” Open-ended responses were scored as correct or incorrect according to a codebook developed by the first author based on the discussion of items by Tullis and Goldstone (2017) and guidance by Tullis (personal communication, September 18, 2017). After six hours of training, the first author and a research assistant coded a random subsample of 50 cases, achieving 100% agreement on one item, 98% agreement with interrater reliabilities of Cohen’s $\kappa = .94$, .66, and .00 on three items, and Cohen’s κ between .70 and .81 on the remaining four items. The two lowest values of κ (.66 and .00) were deemed acceptable because they were obtained with only one disagreement on each item that were rarely scored as correct. After resolving remaining disagreements via discussion, the remainder of the data were divided in half and independently scored by the first author and research assistant. CSCI has been described and discussed in the preceding literature review.

Data Analysis

Prior to analysis, questionnaire pages on which participants spent less than one second per item were replaced with missing values to eliminate poor quality data due to participant speeding (Wood, Harms, Lowman, & DeSimone, 2017). Because measures of systems thinking skills were dichotomous (0 = *incorrect*, 1 = *correct*), they were treated as reflective indicators of latent variables in item response theory (IRT) models (Raykov, 2017) using diagonally weighted least squares (WLSMV) for models with categorical variables in Mplus 8.1 (Muthén & Muthén, 2017). IRT estimates latent abilities—in this case, systems thinking skills—based on responses to categorical items, as well as the likelihood of correct responses to each item as a function of individuals' latent ability level.

For systems thinking skills measures, the plausibility of unidimensionality was tested via confirmatory factor analysis (CFA) and inspection of overall model fit and local fit as indicated by individual items' R^2. If unidimensionality was implausible, the dimensionality of each measure in isolation was explored by comparing competing exploratory factor analysis (EFA) models with different numbers of factors. The number of factors to extract was decided by evaluating Eigenvalues, model Chi-square difference tests, and evaluating the interpretability of rotated loadings. The suitability of one-parameter (i.e., Rasch) IRT models were compared to less parsimonious two-parameter IRT models via model difference tests and comparison of BIC. Whereas one-parameter IRT models fix all items' discriminations (i.e., how sharply they distinguish between individuals of differing abilities), two-parameter IRT models estimate items' discriminations separately. Up to this point, only two-parameter IRT models had been estimated. For each unidimensional IRT model, a subset of the best-performing items reflecting a wide range of abilities—as indicated by item discrimination, item difficulty, item information curves, and test information functions—were retained (Raykov, 2017). Next, these reduced unidimensional systems thinking measures were combined and their dimensionality explored via EFA. See Appendix A for additional statistical details of all analyses, including model comparisons.

Results

For the department store task (DS; Sterman, 2002), a two-factor, one-parameter model containing all four original items was retained (see Table 2). These factors were positively correlated, $r = .43$ ($p < .0001$). The first factor was interpreted as reflecting the ability to read a graph and as such were excluded from subsequent analyses, the second as systems thinking about relationships between stocks and flows. Negative standardized threshold values indicate that the first factor contained relatively easy items, and positive standardized threshold values indicate that the second factor contained relatively difficult items.

Table 2

Two-factor, one-parameter IRT model of the department store task

Factor	Item text	Std. Loading	*SE*	*p*	Std. Threshold	*SE*	*p*
1	During which minute did the most people enter the store?	0.951	0.011	<.0001	-1.144	0.076	<.0001
	During which minute did the most people leave the store?	0.951	0.011	<.0001	-0.941	0.07	<.0001
2	During which minute were the most people	0.951	0.011	<.0001	0.639	0.064	<.0001

in the store?						
During which minute were the fewest people in the store?	0.951	0.011	<.0001	0.639	0.064	<.0001

Note. $\chi^2(4, n = 444) = 4.62$, $p = .3285$, RMSEA = .019 (90% CI: .000, .076), CFI = 1.000, TLI = .999.

For the Groundwater System Dynamic Nature Questionnaire (GDN; Ben-Zvi Assaraf & Orion, 2005), a one-factor, one-parameter model containing five of the initial eight items was retained (see Table 3). This measure was interpreted as recognizing interconnections and relationships between system entities in the context of the hydro-cycle. Negative item difficulties indicate that these items were relatively easy.

Table 3

One-factor, one-parameter IRT model of the Groundwater System Dynamic Nature Questionnaire

Item text	Discrimination	*SE*	*p*	Difficulty	*SE*	*p*
Rocks don’t influence the composition of the water that penetrates them.	0.932	0.078	<.0001	-0.935	0.108	<.0001
Only when rocks are cracked can water penetrate them.	0.932	0.078	<.0001	-0.575	0.096	<.0001
Ground water can be found only in rainy areas.	0.932	0.078	<.0001	-1.457	0.134	<.0001
Some of the wells in the United States contain polluted water.	0.932	0.078	<.0001	-1.403	0.131	<.0001
Rain that falls on the surface and penetrates within the soil can reach a depth of several meters.	0.932	0.078	<.0001	-0.885	0.105	<.0001

Note. $\chi^2(9, n = 443) = 8.479$, $p = .4866$, RMSEA = .000 (90% CI: .000, .051), CFI = 1.000, TLI = 1.002.

For the Cyclic Thinking Questionnaire (CTQ; Ben-Zvi Assaraf & Orion, 2005), a one-factor, two-parameter model containing five of the initial six items was retained (see Table 4).

This measure was interpreted as understanding cycles, feedback, and flows in the context of the hydro-cycle. The range of items' difficulties indicates that the measure identifies a range of abilities.

Table 4

One-factor, two-parameter IRT model of the Cyclic Thinking Questionnaire

Item text	Discrimination	*SE*	*p*	Difficulty	*SE*	*p*
Clouds are the starting point of the water cycle and the tap at home is its end point.	0.682	0.106	<.0001	-0.918	0.152	<.0001
The amount of water in the ocean is growing from day to day because rivers are continually flowing into the ocean.	0.759	0.112	<.0001	-0.719	0.126	<.0001
The increase in evaporation as an effect of the earth's global warming may lead to a decrease in the amount of water on earth.	1.305	0.206	<.0001	0.161	0.076	.034
If the population on earth continues to grow, water consumption will increase, thus decreasing the amount of water in earth.	1.988	0.472	<.0001	0.048	0.067	.476
The amount of water that evaporates into the atmosphere from the entire surface of the earth is not equal to the amount of rain that falls on the earth's surface.	0.843	0.125	<.0001	0.561	0.108	<.0001

Note. $\chi^2(5, n = 443) = 14.869$, $p = .0109$, RMSEA = .067 (90% CI: .029, .107), CFI = .984, TLI = .968.

For the translated Systems Thinking Assessment (STA; K. C. Constantinide, 2015; K. Constantinide et al., 2014), 15 items from the original 29 were selected from the retained 24-item, one-factor, two-parameter model (see Table 5). This measure was interpreted as identifying and recognizing system components, relationships and interconnections, understanding temporal causal dynamics in everyday contexts. The final items were selected to maximize the breadth of the test information function by including high-discrimination items with the largest range of difficulties. Nevertheless, negative item difficulties indicate that these items are relatively easy.

Table 5

One-factor, two-parameter IRT model of the Systems Thinking Assessment

Item text	Discrimination	*SE*	*p*	Difficulty	*SE*	*p*
Which parts are ABSOLUTELY NECESSARY for a bicycle to roll when someone pushes it?	1.018	0.143	<.0001	-1.042	0.119	<.0001
A parent is making their four-year old child a car out of a shoe box. What does the car ABSOLUTELY NEED to have in order to roll, besides the shoe box?	1.426	0.222	<.0001	-1.099	0.105	<.0001
A car is driven on a highway and travels a very long distance. Which of the following statements is WRONG?	1.056	0.159	<.0001	-1.281	0.136	<.0001
A country starts to equip its army with heavy weapons. A hostile country becomes aware of it, and so also begins to arm with heavy weapons. What is most likely to happen in the near future?	1.276	0.184	<.0001	-1.166	0.114	<.0001
This year a shoe company managed to increase their number of customers. They have decided to use this year's income to increase the variety of their products. What is most likely to happen to their number of customers in the next year?	1.002	0.149	<.0001	-1.236	0.136	<.0001
In a remote forest, some trees wither and die. But the total number of the wood's trees remains the same. That means that...	1.290	0.187	<.0001	-1.146	0.111	<.0001
Is there a chance that a dollar bill that I have spent on groceries could end up in my hands?	1.157	0.172	<.0001	-1.267	0.129	<.0001
Winter is coming. In order for a family to decide if their central heating is working properly, it would be SUFFICIENT for them to...	0.860	0.120	<.0001	-0.953	0.127	<.0001
A refrigerator operating correctly is best defined by...	1.352	0.197	<.0001	-1.180	0.110	<.0001
In a certain place, snakes eat mice, and mice eat wheat. What is most likely to happen to the snakes if all of the plants in the area die out?	1.254	0.180	<.0001	-1.109	0.112	<.0001
Water from dams is used for irrigation (watering plants on farms). In order for the dam to always hold at least some water, it is NECESSARY for...	0.964	0.132	<.0001	-0.990	0.119	<.0001
A food chain of animals that live near a river in Pennsylvania is presented below. The diagram shows that bass eat frogs and water snakes, water snakes eat frogs, and frogs eat insects. What will happen to the rest of the animals if the insects near the river go extinct?	0.889	0.121	<.0001	-0.756	0.111	<.0001
Nowadays we create a lot of garbage. Garbage pollutes the environment. What can be done to	1.420	0.245	<.0001	-1.369	0.125	<.0001

reduce environmental pollution from garbage?						
A basketball team is one of the best teams in the NBA this year. Which of the following sentences is WRONG?	1.883	0.381	<.0001	-1.312	0.107	<.0001
A factory that manufactures plastic bottles has started operating for fewer hours than last year. The fact that the factory operates for fewer hours...	1.322	0.204	<.0001	-1.245	0.118	<.0001

Note. $\chi^2(90, n = 457) = 101.199$, $p = .1972$, RMSEA = 0.017 (90% CI: .000, .031), CFI = 0.997, TLI = 0.997.

For the Complex Systems Concepts Inventory (CSCI; Tullis & Goldstone, 2017) a two-factor, two-parameter model containing nine of the original 12 items was retained (see Table 6). These factors were not significantly correlated, $r = .23$ ($p = .076$). The first factor was interpreted as systems thinking about complex systems dynamics such as emergence of macroscopic phenomena from (sometimes hidden) micro-level interactions, the second as making predictions about the long-term consequences of local interactions for a whole system. Mostly positive standardized threshold values indicate that the items on both factors were relatively difficult.

Table 6

Two-factor, two-parameter IRT model of the Complex Systems Concepts Inventory

	Factor		
Item text	1	2	Threshold
The hiker below wants to get to the highest peak on the mountain range. Unfortunately, it is very foggy and he can only see a couple of feet in any direction. He decides to walk in whatever direction will raise him up the highest amount. Why might it be important to add in a bit of randomness to his movements?	**0.789***	0.008	0.389
There is a world inhabited by As and Bs. Wherever there is an A in this world, on the next generation it grows a B below it (if there's not already one there). Wherever there is a B, on the next generation it grows a B to its left and an A below it (if these letters are not there already). What does a world that initially looks like the grid below look like after two generations?	**0.752***	-0.023	0.490
Instead of storing the exact pattern of zebra stripes in zebra DNA, what is a simple rule for how cells interact that could cause stripes to eventually be formed?	**0.744***	-0.041	0.868

There is a world made of black and white squares. Each square has four neighbor squares: one above, one below, one to the left, and one to the right. The squares all change color from one time to the next by the following rule: if a square has more than one black square neighbor, then it will be black. Otherwise, it will be white. All of the squares change at the same time. If the world starts with the pattern in the grid above, what will it look like at the next time step? Click the squares in the grid below to indicate which will be black at the next time step.	**0.700***	0.079	0.735
You are dropping a set of balls through an obstacle course (as shown below). You want all the balls to fall all the way through the obstacles (the black arcs). Why might it be important to add in a bit of random movement to the balls as they fall?	**0.594***	-0.106	-0.024
A pattern of ridges and troughs can be formed when varnish begins to wrinkle and lift off of wood, as shown below. How can this complex pattern occur?	**0.568***	0.234	1.752
Some groups of fireflies will begin to synchronize their flashing after spending some time together in an area. How might large groups synchronize their flashing?	**0.514***	0.077	0.411
There are four kinds of soda in a city: Yaz, Jot, Mup, and Fet. The people in the city are very influenced by each other, and if somebody sees another person drinking a soda, they will then drink the same soda next time. If every person drinks a soda every day in a cafe, but the four soft drinks start off equally popular, then in 3 years, what is the likely outcome?	-0.005	**0.858***	0.574
A large group of children live in a neighborhood. Each child randomly prefers a red, blue, orange, or green toy, so that these colors are equally preferred across the neighborhood. The children are constantly moving around the neighborhood and playing with other kids. As they randomly move about, they look to see the preferred color of the most other kids around them. They switch their toy preference to the one preferred by the most children that they see at any moment. What will happen to toy preferences over time?	0.240	**0.625***	0.162
Eigenvalue	3.993	1.401	

Note. Factor loadings are standardized geomin-rotated loadings. Primary loadings are bolded. *p < .05. $\chi^2(19, n = 461) = 24.592$, $p = .1744$, RMSEA = 0.025 (90% CI: .000, .051), CFI = 0.992, TLI = 0.984.

To investigate the dimensionality of the preceding measures of systems thinking skills, a six-factor CFA was estimated, which indicated an unsatisfactory fit of the model to the data, $\chi^2(584, n = 463) = 759.509$, $p < .0001$, RMSEA = 0.025 (90% CI: .020, .030), CFI = 0.976, TLI = 0.975. Therefore, an EFA was estimated using oblique geomin rotation to explore the dimensionality of these items in combination after dropping one STA and two CSCI items for having empty cells in a bivariate table with two or more other items. Identifying the number of factors to retain in EFAs with dichotomous indicators is an active area of research, with suggestions including retaining the model with the fewest factors with RMSEA < .04 (Barendse, Oort, & Timmerman, 2015) or the model with the fewest factors with CFI > .95 and TLI > .95 (Clark & Bowles, 2018). Because these guidelines ambiguously suggested retaining either a two- or four-factor model, Clark and Bowles's (2018) program using the MplusAutomation package (Hallquist & Wiley, 2018) for R to run simulations in Mplus was used to conduct a series of simulations to examine the performance of these and other fit indices for retaining an EFA model based on this dataset. For each of the 1–5 factor solutions, 1000 simulated datasets were generated using estimated model parameters and population parameters. Next, EFAs were conducted on these six datasets (see Appendix B). Extant recommendations for using RMSEA, CFI, and TLI appeared to underfactor in comparison to the factor structure underlying the simulated data (see Appendix B). Instead, the incremental decrease of RMSEA between at least .005 and .010, incremental increase of CFI between at least .005 and .010, and incremental increase of TLI between at least .005 and .015 most consistently identified the sample factor structures matching the population structures of the underlying simulated data (see Appendix B). Applied to the EFA of all systems measures of systems thinking skills, incremental improvement of fit indices suggested retaining a four- or five-factor solution (see Appendix A). Because the five-factor solution contained a factor with only weak loadings, the four-factor model was retained, $\chi^2(402, n = 463) = 519.925$, $p < .0001$, RMSEA = .025 (90% CI: .018, .031), CFI = 0.983, TLI = .977, SRMR = 0.062. (See Appendix C for factor loadings and factor correlations.)

These items in combination comprised the Multidimensional Inventory of Systems Thinking Skills (MISTS). Although each factor's strongest loadings were comprised of items from single measures, all measures except DS had items that exhibited substantial cross-loading. Therefore, the poor fit of the initial CFA appears to have been attributable primarily to these cross-loadings, as opposed to more dramatic model misspecification. Although cross-loadings suggest an underlying factor structure not attributable to constructs unique to each measure, neither of the potential hypothesized factor structures (see Table 1) appear to be clearly supported.

Discussion

This study examined the potential for existing measures of systems thinking skills to be shortened and combined into the Multidimensional Inventory of Systems Thinking Skills (MISTS) to measure a range of systems thinking skills and abilities. The dimensionality of the department store task (DS; Sterman, 2002), the Groundwater System Dynamic Nature Questionnaire (GDN; Ben-Zvi Assaraf & Orion, 2005), the Cyclic Thinking Questionnaire (CTQ; Ben-Zvi Assaraf & Orion, 2005), the translated Systems Thinking Assessment (STA; K. C. Constantinide, 2015; K. Constantinide et al., 2014), and the Complex Systems Concepts Inventory (CSCI; Tullis & Goldstone, 2017) were tested and explored. Unidimensional factors with items with high information functions across a range of difficulties were combined into the MISTS. It was predicted that the factor structure of the MISTS would reflect typologies of systems thinking skills in the literature (Ben-Zvi Assaraf & Orion, 2005; K. Stave & Hopper, 2007). Contrary to this prediction, the five-factor structure of the MISTS appeared to reflect the measures from which items originated, rather than theorized typologies of systems thinking skills. These MISTS factors had unique patterns of correlations with some—but not all—measures of endorsing systems concepts. Overall, these results suggest that the MISTS may hold promise as a valid instrument for measuring systems thinking skills.

It was proposed that in order to maximize the utility of a measure of systems thinking skill, it should: (a) reflect a wide range of systems thinking skills, (b) capture a wide range of systems thinking abilities, (c) not require knowledge of a particular content area—i.e., be domain-independent, (d) be easy to administer, (e) be easy to score, and (f) be as short possible so as to minimize burden on participants while satisfying the preceding criteria. The MISTS appears to have met these criteria. The MISTS's five-factor structure, in which all factors were significantly correlated with at least two other factors (see Appendix B), suggests that it reflects multiple distinct—but related—skills. Additionally, the range of item difficulties indicates that it captures a breadth of ability levels. Although it could be argued that DS relies on graph reading ability and that GDN and CTQ reflect understandings of the hydro-cycle, the weak correlation between the two DS factors and the non-significant correlation between GDN and CTQ suggest that they are not intimately related. The MISTS can be administered as a questionnaire, and the closed-ended items are easy to score. At 29 items, the MISTS is shorter than the initial set of systems thinking measures.

It was expected that the MISTS dimensions would be associated with individuals' endorsement of systems concepts, indicating a degree of convergent validity. Given the difference between endorsing systems concepts and systems thinking skills per se, however, it was expected that these associations would be modest, indicating a degree of discriminant validity. All of the MISTS factors were significantly correlated with at least two measures of endorsing systems concepts, in support of the MISTS's convergent validity. With the exception of the GDN- and CTQ-dominated factors of the MISTS, the magnitude of these partial correlations was modest ($r < .35$), suggesting their discriminant validity from merely endorsing systems concepts. The GDN and CTQ were substantially correlated ($r > .60$) with one and three measures of endorsing systems concepts, respectively, suggesting that these two factors may have reflected—at least in part—the application of simple systems heuristics (e.g., things are related to each other, systems are stable) rather than true systems thinking ability. As such, more validation of the MISTS is warranted.

Limitations and Future Research

The MISTS contained items of primarily low difficulty, suggesting that it is best at discriminating between individuals of relatively low systems thinking ability. As such, the MISTS may be limited in its utility for capturing higher abilities of systems thinking. Given the established difficulty of systems thinking (Booth Sweeney & Sterman, 2007; Cronin et al., 2009; Kapmeier et al., 2017), however, measuring low-level systems thinking is still of value. Future research should include additional, more challenging tests of systems thinking, such as closed-ended versions of other stock-and-flow tasks (e.g., Booth Sweeney & Sterman, 2000).

This survey contained a large number of items, and the systems thinking skills measures were presented to participants after the measures of endorsing systems concepts. As such, participants may have experienced cognitive fatigue by the end of the survey. If so, items may have functioned differently than they would have if participants had been less fatigued. To mitigate this threat to construct validity, responses that were likely answered uncaringly were replaced with missing values. Future studies should test whether the MISTS maintains the same factorial structure when administering it in a shorter survey.

In addition to confirming the factor structure of the MISTS, it is important to establish its predictive validity and to further support its convergent and discriminant validity. In particular, if the MISTS is a valid test of systems thinking skills, it should be able to predict certain outcomes (e.g., performance on systems dynamics tasks) and be predicted by certain experiences (e.g., taking a class in complex systems, playing complex boardgames). In terms of convergent and discriminant validity, it would be important to explore the components of intelligence to which systems thinking are related, and to simultaneously establish that performance on the MISTS cannot be reduced to general intelligence.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Addison Wesley Longman.

Baghaei Lakeh, A., & Ghaffarzadegan, N. (2016). The dual-process theory and understanding of stocks and flows. *System Dynamics Review*, *32*(3–4), 309–331. doi: 10.1002/sdr.1566

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 87–101. doi: 10.1080/10705511.2014.934850

Ben-Zvi Assaraf, O., Dodick, J., & Tripto, J. (2013). High school students' understanding of the human body system. *Research in Science Education*, *43*(1), 33–56. doi: 10.1007/s11165-011-9245-2

Ben-Zvi Assaraf, O., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching*, *42*(5), 518–560. doi: 10.1002/tea.20061

Ben-Zvi Assaraf, O., & Orion, N. (2010a). Four case studies, six years later: Developing system thinking skills in junior high school and sustaining them over time. *Journal of Research in Science Teaching*, *47*(10), 1253–1280. doi: 10.1002/tea.20383

Ben-Zvi Assaraf, O., & Orion, N. (2010b). System thinking skills at the elementary school level. *Journal of Research in Science Teaching*, *47*(5), 540–563. doi: 10.1002/tea.20351

Bertalanffy, L. von. (2015). *General system theory: Foundations, development, applications* (Revised edition). New York, NY: George Braziller.

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

Booth Sweeney, L., & Sterman, J. D. (2007). Thinking about systems: Student and teacher conceptions of natural and social systems. *System Dynamics Review*, *23*(2–3), 285–311. doi: 10.1002/sdr.366

Brandstädter, K., Harms, U., & Großschedl, J. (2012). Assessing system thinking through different concept-mapping practices. *International Journal of Science Education*, *34*(14), 2147–2170. doi: 10.1080/09500693.2012.716549

Checkland, P. (1999). *Systems thinking, systems practice: Includes a 30-year retrospective*. Retrieved from http://eprints.lancs.ac.uk/id/eprint/47648

Clark, D. A., & Bowles, R. P. (2018). Model fit and item factor analysis: Overfactoring, underfactoring, and a program to guide interpretation. *Multivariate Behavioral Research*, *53*(4), 544–558. doi: 10.1080/00273171.2018.1461058

Constantinide, K. C. (2015). *Development and validation of a systems thinking assessment instrument for 10-14 year olds* (Doctoral dissertation, University of Cyprus). Retrieved from http://lekythos.library.ucy.ac.cy/handle/10797/14389

Constantinide, K., Michaelides, M., & Constantinou, C. P. (2014). Development of an instrument to measure children's systems thinking. In C. P. Constantinou, N. Papadouris, & A. Hadjigeorgiou (Eds.), *E-Book Proceedings of the ESERA 2013 Conference: Science Education Research For Evidence-based Teaching and Coherence in Learning* (Vol. 11, pp. 13–23). Retrieved from https://www.esera.org/publications/esera-conference-proceedings/esera-2013

Cronin, M. A., Gonzalez, C., & Sterman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes*, *108*(1), 116–130. doi: 10.1016/j.obhdp.2008.03.003

Evagorou, M., Korfiatis, K., Nicolaou, C., & Constantinou, C. (2009). An investigation of the potential of interactive simulations for developing system thinking skills in elementary school: A case study with fifth-graders and sixth-graders. *International Journal of Science Education*, *31*(5), 655–674. doi: 10.1080/09500690701749313

Fischer, H., & Degen, C. (2012). Stock-flow failure can be explained by the task format. In E. Husemann & D. Lane (Eds.), *Proceedings of the 30th International Conference of the System Dynamics Society*. Retrieved from http://www.systemdynamics.org/conferences/2012/index.html

Fischer, H., Kapmeier, F., Tabacaru, M., & Kopainsky, B. (2015, July). *The more you see the less you "get": On the importance of a higher-level perspective for understanding dynamic systems*. Presented at the 33rd International Conference of the System Dynamics Society, Cambridge, MA. Retrieved from http://www.systemdynamics.org/conferences/2015/index.html

Forrester, J. W. (1994). System dynamics, systems thinking, and soft OR. *System Dynamics Review*, *10*(2–3), 245–256. doi: 10.1002/sdr.4260100211

Garson, G. D. (2009). Computerized simulation in the social sciences: A survey and evaluation. *Simulation & Gaming*, *40*(2), 267–279. doi: 10.1177/1046878108322225

Google scholar. (n.d.). Retrieved November 26, 2017, from https://scholar.google.com/scholar?cites=13443129322880353132&as_sdt=5,49&sciodt=0,49&hl=en

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. doi: 10.1080/10705511.2017.1402334

Hopper, M., & Stave, K. A. (2008). Assessing the effectiveness of systems thinking interventions in the classroom. *Proceedings of the 26th International Conference of the System Dynamics Society*. Presented at the Athens, Greece. Retrieved from http://www.systemdynamics.org/conferences/2008/proceed/index.htm

Kainz, D., & Ossimitz, G. (2002). *Can students learn stock-flow-thinking? An empirical investigation*. Presented at the 2002 System Dynamics Conference, Palermo, Italy.

Kapmeier, F., Happach, R. M., & Tilebein, M. (2017). Bathtub dynamics revisited: An examination of déformation professionelle in higher education. *Systems Research and Behavioral Science*, *34*(3), 227–249. doi: 10.1002/sres.2407

Kasperidus, H. D., Langfelder, H., & Biber, P. (2006). Comparing systems thinking inventory task performance in German classrooms at high school and university level. *The 24th International Conference of the System Dynamics Society*. Presented at the Nijmegen, Netherlands. Nijmegen, Netherlands.

Kaufman, G., & Flanagan, M. (2016). Playing the system: Comparing the efficacy and impact of digital and non-digital versions of a collaborative strategy game. *Proceedings of the First International Joint Conference of DiGRA and FDG*. Retrieved from http://www.digra.org/digital-library/forums/13-digra-fdg2106/

Lee, T. D., Jones, M. G., & Chesnutt, K. (2017). Teaching systems thinking in the context of the water cycle. *Research in Science Education*, Advance online publication. doi: 10.1007/s11165-017-9613-7

Levine, R. L., & Doyle, J. K. (2002). Modeling generic structures and patterns in social psychology. In P. I. Davidsen, E. Mollona, V. G. Diker, R. S. Langer, & J. I. Rowe (Eds.), *Proceedings of the 20th International Conference of the System Dynamics Society*. Retrieved from http://www.systemdynamics.org/conferences/2002/index.html

Meadows, D. H. (2008). *Thinking in systems: A primer* (D. Wright, Ed.). White River Junction, VT: Chelsea Green Publishing.

Mitchell, M. (2009). *Complexity: A guided tour*. New York, NY: Oxford University Press.

Muthén, L. K., & Muthén, B. (2017). *Mplus user's guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

National Research Council. (2012). *A framework for k-12 science education: Practices, crosscutting concepts, and core ideas*. Retrieved from http://www.nap.edu/catalog/13165/a-framework-for-k-12-science-education-practices-crosscutting-concepts

Ossimitz, G. (2002). *Stock–flow thinking and reading stock–flow-related graphs: An empirical investigation in dynamic thinking abilities*. Presented at the 2002 System Dynamics Conference, Palermo, Italy.

Pala, Ö., & Vennix, J. A. M. (2005). Effect of system dynamics education on systems thinking inventory task performance. *System Dynamics Review*, *21*(2), 147–172. doi: 10.1002/sdr.310

Raykov, T. (2017). *Item response theory*. Philadelphia, PA: Course booklet. Statistical Horizons.

Riess, W., & Mischo, C. (2010). Promoting systems thinking through biology lessons. *International Journal of Science Education*, *32*(6), 705–725. doi: 10.1080/09500690902769946

Senge, P. M. (2006). *The fifth discipline: The art and practice of the learning organization*. New York, NY: Doubleday/Currency.

Stave, K. A., Beck, A., & Galvan, C. (2015). Improving learners' understanding of environmental accumulations through simulation. *Simulation & Gaming*, *46*, 293–321. doi: 10.1177/1046878114531764

Stave, K., & Hopper, M. (2007). What constitutes systems thinking? A proposed taxonomy. In J. Sterman, R. Oliva, R. S. Langer, J. I. Rowe, & J. M. Yanni (Eds.), *Proceedings of the 25th International Conference of the System Dynamics Society*. Retrieved from http://www.systemdynamics.org/conferences/2007/index.htm

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

Sterman, J. D. (2010). Does formal system dynamics training improve people's understanding of accumulation? *System Dynamics Review*, *26*(4), 316–334. doi: 10.1002/sdr.447

Tullis, J. G. (2017, September 18). *Re: Complex systems concepts inventory (Tullis and Goldstone, 2017)*.

Tullis, J. G., & Goldstone, R. L. (2017). Instruction in computer modeling can support broad application of complex systems knowledge. *Frontiers in Education*, *2*. doi: 10.3389/feduc.2017.00004

Weinhardt, J. M., Hendijani, R., Harman, J. L., Steel, P., & Gonzalez, C. (2015). How analytic reasoning style and global thinking relate to understanding stocks and flows. *Journal of Operations Management*, *39–40*, 23–30. doi: 10.1016/j.jom.2015.07.003

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454–464. doi: 10.1177/1948550617703168

Appendix A: Model Fit Statistics and Comparisons

See accompanying Excel document.

Appendix B: Simulation Studies

Appendix C: Exploratory Factor Analysis Results of the Multidimensional Inventory of Systems Thinking Skills

Table C1

Exploratory Factor Analysis Results of the Multidimensional Inventory of Systems Thinking Skills

Initial Scale	Item Text	Factor 1	2	3	4	5	6	Threshold
DS	During which minute were the most people in the store?	**0.878***	0.056	-0.096	0.025	0.049	0.003	0.639
DS	During which minute were the fewest people in the store?	**0.987***	0.001	0.027	-0.058	-0.005	-0.008	0.639
GDN	Rocks don't influence the composition of the water that penetrates them.	-0.100	0.269	**0.506***	0.048	0.014	-0.006	-0.638
GDN	Only when rocks are cracked can water penetrate them.	0.110	0.272	**0.367***	0.184	-0.139	0.080	-0.392
GDN	Ground water can be found only in rainy areas.	-0.006	0.280	**0.445***	**0.271***	0.065	0.010	-0.993
GDN	Some of the wells in the United States contain polluted water.	0.047	**0.380***	**0.492***	-0.066	-0.092	-0.053	-0.957
GDN	Rain that falls on the surface and penetrates within the soil can reach a depth of several meters.	-0.019	0.025	**0.768***	-0.055	0.109	-0.045	-0.603
CTQ	Clouds are the starting point of the water cycle and the tap at home is its end point.	-0.147	**0.484***	0.028	**0.337***	0.139	0.037	-0.517
CTQ	The amount of water in the ocean is growing from day to day because rivers are continually flowing into the ocean.	-0.046	0.221	0.146	**0.454***	0.062	-0.012	-0.435
CTQ	The increase in evaporation as an effect of the earth's global warming may lead to a decrease in the amount of water on earth.	0.162	-0.015	-0.079	**0.746***	0.035	-0.030	0.128
CTQ	If the population on earth continues to grow, water consumption will increase, thus decreasing the amount of water in earth.	0.031	0.097	-0.041	**0.845***	-0.005	0.202*	0.042
CTQ	The amount of water that evaporates into the atmosphere from the entire surface of the earth is not equal to the amount of rain that falls on the earth's surface.	**0.369***	-0.229	0.109	**0.573***	-0.005	-0.094	0.361
STA	Which parts are ABSOLUTELY NECESSARY for a bicycle to roll when someone pushes it?	0.196	**0.451***	0.073	0.146	0.154	-0.041	-0.744

STA	A parent is making their four-year old child a car out of a shoe box. What does the car ABSOLUTELY NEED to have in order to roll, besides the shoe box?	**0.309***	**0.551***	0.127	0.072	0.017	0.050	-0.899
STA	A car is driven on a highway and travels a very long distance. Which of the following statements is WRONG?	0.002	**0.672***	0.111	-0.138	0.057	0.106	-0.930
STA	A country starts to equip its army with heavy weapons. A hostile country becomes aware of it, and so also begins to arm with heavy weapons. What is most likely to happen in the near future?	0.031	**0.620***	0.122	-0.090	0.270*	0.009	-0.918
STA	This year a shoe company managed to increase their number of customers. They have decided to use this year's income to increase the variety of their products. What is most likely to happen to their number of customers in the next year?	0.009	**0.752***	0.000	0.056	-0.082	-0.107	-0.875
STA	In a remote forest, some trees wither and die. But the total number of the wood's trees remains the same. That means that...	0.045	**0.500***	0.227	0.075	0.260*	-0.006	-0.906
STA	Winter is coming. In order for a family to decide if their central heating is working properly, it would be SUFFICIENT for them to...	0.081	**0.596***	-0.143	-0.038	0.174	0.031	-0.622
STA	A refrigerator operating correctly is best defined by...	0.191	**0.625***	0.154	0.076	-0.047	0.127	-0.949
STA	In a certain place, snakes eat mice, and mice eat wheat. What is most likely to happen to the snakes if all of the plants in the area die out?	-0.086	**0.603***	-0.064	-0.013	**0.566***	-0.116	-0.867
STA	Water from dams is used for irrigation (watering plants on farms). In order for the dam to always hold at least some water, it is NECESSARY for...	-0.093	**0.832***	-0.133	-0.020	-0.012	-0.031	-0.687
STA	A food chain of animals that live near a river in Pennsylvania is presented below. The diagram shows that bass eat frogs and water snakes, water snakes eat frogs, and frogs eat insects. What will happen to the rest of the animals if the insects near the river go extinct?	0.161	**0.589***	-0.136	-0.078	0.210	-0.159	-0.502
STA	Nowadays we create a lot of garbage. Garbage pollutes the environment. What can be done to reduce environmental pollution from garbage?	-0.004	**0.807***	0.091	0.073	-0.005	-0.216*	-1.119

STA	A basketball team is one of the best teams in the NBA this year. Which of the following sentences is WRONG?	0.114	**0.810***	0.070	-0.060	-0.006	0.070	-1.158
STA	A factory that manufactures plastic bottles has started operating for fewer hours than last year. The fact that the factory operates for fewer hours...	-0.046	**0.801***	0.007	0.139	-0.059	0.030	-0.993
CSCI	There is a world inhabited by As and Bs. Wherever there is an A in this world, on the next generation it grows a B below it (if there's not already one there). Wherever there is a B, on the next generation it grows a B to its left and an A below it (if these letters are not there already). What does a world that initially looks like the grid below look like after two generations? Type the As and Bs in the grid below.	0.038	-0.032	0.120	0.125	**0.879***	0.044	0.490
CSCI	A large group of children live in a neighborhood. Each child randomly prefers a red, blue, orange, or green toy, so that these colors are equally preferred across the neighborhood. The children are constantly moving around the neighborhood and playing with other kids. As they randomly move about, they look to see the preferred color of the most other kids around them. They switch their toy preference to the one preferred by the most children that they see at any moment. What will happen to toy preferences over time?	0.032	0.218	-0.080	0.004	0.009	**0.726***	0.162
CSCI	You are dropping a set of balls through an obstacle course (as shown below). You want all the balls to fall all the way through the obstacles (the black arcs). Why might it be important to add in a bit of random movement to the balls as they fall?	0.104	0.078	0.233	-0.008	**0.403***	-0.001	-0.024
CSCI	Some groups of fireflies will begin to synchronize their flashing after spending some time together in an area. How might large groups synchronize their flashing?	**0.329***	0.047	0.093	0.044	**0.256***	0.102	0.411

CSCI	There is a world made of black and white squares. Each square has four neighbor squares: one above, one below, one to the left, and one to the right. The squares all change color from one time to the next by the following rule: if a square has more than one black square neighbor, then it will be black. Otherwise, it will be white. All of the squares change at the same time. If the world starts with the pattern in the grid above, what will it look like at the next time step? Click the squares in the grid below to indicate which will be black at the next time step.	**0.271***	**0.287***	-0.118	0.125	**0.281***	0.134	0.735
CSCI	There are four kinds of soda in a city: Yaz, Jot, Mup, and Fet. The people in the city are very influenced by each other, and if somebody sees another person drinking a soda, they will then drink the same soda next time. If every person drinks a soda every day in a cafe, but the four soft drinks start off equally popular, then in 3 years, what is the likely outcome?	-0.027	-0.066	0.056	-0.043	0.023	**0.788***	0.574
CSCI	The hiker below wants to get to the highest peak on the mountain range. Unfortunately, it is very foggy and he can only see a couple of feet in any direction. He decides to walk in whatever direction will raise him up the highest amount. Why might it be important to add in a bit of randomness to his movements?	**0.395***	0.090	-0.019	0.054	**0.345***	0.076	0.389

Note. Bolded standardized loadings indicate primary factor loadings. *Loading significant at $p <$ .05.

Table C2

Partial Correlations of Factors of the Multidimensional Inventory of Systems Thinking Skills

Factor	1	2	3	4	5
1					
2	0.459*				
3	0.181	0.428*			
4	0.381*	0.345*	0.241		

5	0.378*	0.434*	0.088	0.332*	
6	0.250*	0.200	-0.001	0.079	0.162

*$p < .05$.