Journal of Science Education and Technology
https://doi.org/10.1007/s10956-023-10027-2

Comparing Self-Report Assessments and Scenario-Based Assessments of Systems Thinking Competence

Kirsten A. Davis[1] · Dustin Grote[2] · Hesam Mahmoudi[3] · Logan Perry[4] · Navid Ghaffarzadegan[5] · Jacob Grohs[6] · Niyousha Hosseinichimeh[5] · David B. Knight[6] · Konstantinos Triantis[5]

Accepted: 28 December 2022
© The Author(s) 2023

Abstract

Self-report assessments are used frequently in higher education to assess a variety of constructs, including attitudes, opinions, knowledge, and competence. Systems thinking is an example of one competence often measured using self-report assessments where individuals answer several questions about their perceptions of their own skills, habits, or daily decisions. In this study, we define systems thinking as the ability to see the world as a complex interconnected system where different parts can influence each other, and the interrelationships determine system outcomes. An alternative, less-common, assessment approach is to measure skills directly by providing a scenario about an unstructured problem and evaluating respondents' judgment or analysis of the scenario (scenario-based assessment). This study explored the relationships between engineering students' performance on self-report assessments and scenario-based assessments of systems thinking, finding that there were no significant relationships between the two assessment techniques. These results suggest that there may be limitations to using self-report assessments as a method to assess systems thinking and other competencies in educational research and evaluation, which could be addressed by incorporating alternative formats for assessing competence. Future work should explore these findings further and support the development of alternative assessment approaches.

Keywords Self-report assessments · Scenario-based assessment · Systems thinking · Competence assessment

Introduction

An important yet nebulous aspiration of higher education involves cultivating future leaders for a complex and unknown tomorrow. Writing about the future, Paul noted, "Governmental, economic, social, and environmental problems will become increasingly complex and interdependent… The forces to be understood and controlled will be corporate, national, trans-national, cultural, religious, economic, and environmental, all intricately intertwined" (1993, p.13). More recently, Wheatley depicted the world as an "interconnected planet of uncertainty and volatility" where changes in one area can dramatically and surprisingly impact changes in an interconnected area (2005, p. 114). As problems continue to grow in complexity and connectivity, individuals in search of innovative ideas and solutions are being asked to exhibit systems thinking capabilities characterized by an ability to see the world as a complex interconnected system where different parts can influence each other and the interrelationships determine system outcomes (Senge, 2006; Sterman, 2000).

To prepare for such complex challenges, governmental bodies, industry, and funding agencies have all pressed colleges and universities to offer experiences that prepare students to tackle broad problems. Students should be able to integrate and connect ideas and perspectives across multiple

✉ Kirsten A. Davis
kad@purdue.edu

David B. Knight
dbknight@vt.edu

1 School of Engineering Education, Purdue University, West Lafayette, IN, USA

2 Department of Teacher Education, Weber State University, Ogden, UT, USA

3 Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA

4 Department of Civil Engineering, University of Nebraska, Lincoln, NE, USA

5 Department of Industrial and Systems Engineering, Virginia Tech, Falls Church, VA, USA

6 Department of Engineering Education, Virginia Tech, Blacksburg, VA, USA

Published online: 21 March 2023

disciplines and content areas to discover and invent new solutions (e.g., National Academy of Engineering, 2004; National Academy of Sciences, 2004; National Institutes of Health, 2006; National Research Council, 2012). The National Research Council's (2012) *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century* identified "systems thinking" as an area of strong overlap between discipline-based standards and deeper learning/twenty-first century skills (i.e., knowledge or skills that can be transferred or applied in new situations). Systems thinking has also been identified as a broader core competency in sustainability research and problem solving (Wiek et al., 2011) and sustainability literacy (ACPA, 2008; Connell et al., 2012; Dale & Newman, 2005; Svanström et al., 2008).

Despite widespread consensus around the importance of this competency for graduates, colleges and universities have several challenges to overcome to support its development. First, discipline-based organizational structures that typically organize curricula (Warburton, 2003) can hinder student development of systems thinking and interdisciplinary problem solving skills (Svanström et al., 2008; Warburton, 2003). Second, systems thinking is a challenging competency to assess, and in the absence of other validated measures, programs have traditionally relied on students' self-assessments as a form of evidence. Our paper addresses this latter challenge and problematizes the reliance on self-report assessments for systems thinking. We present comparisons of engineering students' self-assessments of systems thinking and several related competencies (i.e., critical thinking, interdisciplinary skills, contextual competence) and their performance on two newly developed scenario-based assessments of systems thinking to address the following research question: Are students' scores on scenario-based assessments of systems thinking related to their scores on self-report assessments of related competencies? Our findings demonstrate that the self-assessments do, indeed, relate to one another, but we do not see a relationship with performance on either scenario-based assessment. These results raise important questions about what information is obtained through each assessment method.

Self-Report Assessments

Self-report assessments are frequently used in educational research and assessment, particularly in the context of co-curricular student activities (e.g., service learning, study abroad) and national surveys of student outcomes (e.g., National Survey of Student Engagement, NSSE). Prior research on self-report assessments has argued both for and against their use in educational research and assessment (Bowman & Hill, 2011; Chan, 2009; Miller, 2012; Porter, 2011). Many of these arguments are based on surveys that ask students to self-report learning gains (i.e., how much they have learned between time A and time B) or changes in attitudes over time. For example, Bowman and Hill (2011) explored social desirability bias (i.e., tendency of students to respond in a way that would be viewed favorably by others) and halo effect (i.e., where a general positive or negative impression of an experience influences responses on specific items) in NSSE respondents and found that these issues were significant in first-year college students, but negligible in later years. Similarly, studies of the Wabash data set, a national data set that combines institutional data (e.g., enrollment and test scores) with student survey data, revealed differences between student self-reported gains and longitudinal measures of development, including biases that varied across institutional type and student characteristics (Bowman, 2010, 2011). Based on these results, Bowman (2011) argues that students with more reason to reflect on their learning (e.g., if they are concerned about their performance) may report their learning gains more accurately. Bowman (2011) suggests that using self-reported learning gains in research is questionable but may be overcome by accounting for known biases.

Porter (2011) takes a stronger stand, arguing that self-reported learning gains are not valid, using NSSE as an example. Building on research on instrument development, memory, and recall, Porter claims that students cannot be expected to estimate how their learning has changed over time accurately. In other studies, Porter has shown that students struggle to report more concrete pieces of information accurately, such as books used in courses and performance in courses (Porter, 2011; Rosen et al., 2017). Further, Porter (2013) has suggested a theory to explain student approaches in responding to survey questions asking them to self-report learning gains. This theory builds on the idea that students follow a complex seven-step process when trying to answer such a question accurately, and thus, students may actually treat these questions as attitudinal rather than factual, (i.e., "how do I feel about my learning" rather than objectively assessing learning gains). Although these perspectives provide helpful insights into the challenges of using self-report assessments to measure student learning gains (i.e., change over time), measuring gains over time is not the same as measuring attitudes or competence at a particular moment in time, which is the focus of the current study.

Taking a broader view of self-report surveys beyond those focused on learning gains, Chan (2009) suggests that the traditional concerns about self-report assessments are too broad (e.g., social desirability bias, common methods variance)—asserting that these critiques are valid in specific cases but not for every study. In many cases, self-report assessments may be as valid as other measures and, occasionally, may be even more so (e.g., when studying individual perspectives or attitudes; Chan, 2009). Pike (2011) presents a more nuanced

view of self-report assessments, suggesting that they can be used effectively in educational research when they are supported by intentional use of theory in their design and when interpreting results. Most validation studies of self-report measures focus on assessing criterion validity; that is, they compare the self-report results to other outcomes that are expected to align with the construct being measured. In some cases, this expected alignment is clear (e.g., self-reported learning in a course could be compared to final grades). However, when the construct of interest is a more abstract concept (e.g., critical thinking), theory is necessary to support the researcher's selection of the comparison measures that are used in the validation study (Pike, 2011).

Research on self-assessment of competence suggests that experts can more accurately assess their competence level than novices (Kruger & Dunning, 1999). Several explanations for this effect have been suggested, including (but not limited to) the following: (a) greater expertise produces enhanced metacognitive skills, which allow people to judge their level of competence more accurately (Dunning & Kruger, 2002; Ehrlinger et al., 2008; Kruger & Dunning, 1999); (b) different perceptions of difficulty in a task lead to different assessments (Burson et al., 2006); and (c) lacking self-confidence in a task results in more arbitrary assessments (Händel & Dresel, 2018). This self-assessment pattern has been identified across a variety of domains and populations, including college students assessing their performance on exams (Händel & Dresel, 2018). Because students are novices in their fields, they may be vulnerable to inaccurate self-assessment, especially in competence areas that could be expected to develop over the course of a career. Finding alternatives to students' self-assessment of such competences is therefore particularly important. Indeed, initial studies of self-report surveys of students' competence have revealed biases in their results. For example, Anderson et al. (2017) compared self-reports to situational judgment tests (SJTs, whereby respondents are provided a problem or scenario and asked how they would respond) and discrete-choice experiments (DCEs, whereby respondents are asked to state their preferences or choices across a set of options), finding that both SJTs and DCEs may mitigate some of the bias issues in self-report surveys assessing interpersonal and intrapersonal skills.

The purpose of our study is to compare the results of self-report assessments and scenario-based assessments of a more abstract concept, systems thinking. In alignment with Pike's (2011) suggestions, we provide theoretical support for the scenario-based assessments we use. Through our analysis, we make assertions about the validity of self-report surveys for assessing this competency and contribute to the ongoing discussion about when and how to use self-report assessments in educational research and assessment.

Theoretical Perspectives Framing Systems Thinking Assessments

Systems thinking is the ability to see the world as a complex interconnected system where different parts can influence each other and the interrelationships determine system outcomes (Senge, 2006; Sterman, 2000). Such a perspective results in seeing multiple stakeholders, focusing on trends rather than single events, and considering unintended consequences of actions intended to improve a system outcome. Because systems thinking is a concept that has been developed across disciplines, there are a variety of definitions and conceptualizations (Mahmoudi et al., 2019). Our own orientation as systems thinking researchers is that we embrace the methodological pluralism that gives rise to myriad definitions and tools. We believe not only that this diversity of approaches is beneficial but also that it is to be expected largely because of the nature of the problems systems thinking aims to solve—namely, ill-structured and so-called wicked problems. Such problems are often characterized by their ambiguity and socio-technical complexity, where no clear single solution exists and where problem-solvers must make sense out of insufficient and/or overwhelming extensive information to scope out and implement solutions which may in turn give rise to new problems.

Aligned with this framing, we believe systems thinking to be a metacognitive competency, marked by the ability to critically and flexibly reason through complexity and multiple dimensions in any decision-making or problem-solving context. Although we believe that some skills transfer across domains within systems thinking, specific knowledge domains (e.g., systems engineering field) still warrant their own catered definitions and approaches. Further, we acknowledge that this methodological pluralism leads to significant and, at times, confusing discrepancies between definitions of systems thinking as well as overlap with other types of thinking discussed in literature such as critical thinking, creativity, and design thinking. Our understanding of systems thinking as theoretically related to these other constructs informed our selection of the assessment tools that we used, such as assessments of critical thinking and contextual competence. Although these assessments do not all use the terminology of systems thinking, the construct definitions and items indicate similar concepts.

Considering the varied definitions of systems thinking and associated tools, a recent systematic literature review by Dugan et al. (2022) mapped out a range of assessments being used in engineering education, identified 27 assessments in total, and categorized both assessment types and formats. A majority of the assessments (19/27) identified were behavior-based, which involve assessing knowledge or skills from responses or artifacts of the participant, while the other

types included preference-based, self-reported, and cognitive activation. The specific formats of the assessments included mapping formats (e.g., concept mapping), scenario-based responses, oral, fill-in-the-blank, multiple-choice, virtual reality, cognitive activation via functional near-infrared spectroscopy (fNIRS), or open-ended in format (i.e., participant responses are not based on prepopulated language or options).

In this paper, we employ two behavior-focused, scenario-based assessments which measure different aspects of students' understanding of systems and then compare those measures to the several existing self-reported assessments of systems thinking and other competencies which literature suggests are related. Each of these scenario-based assessments is based on a theoretical framework of systems thinking, as described in the following sections. As outlined above, we believe in methodological pluralism and that there is great value in a wide range of assessments for systems thinking. However, of the types and formats identified by Dugan et al. (2022), we find value in interrogating the relationships between behavior-based and self-reported assessments given that these two modalities most naturally mimic the teaching and learning environments of university education. Specifically, courses regularly use behavior or performance-based metrics to assess learning, and co-curricular and/or program assessment in universities often use self-reported assessments as a quick, convenient way to try to capture pre/post change in self-reported attitudes, values, beliefs, or behaviors. Further, in a comparison study of cognitive activation and self-reported assessments, Hu and Shealy (2018) found no correlation and highlighted that further efforts to study relationships between self-reported assessments and other types of assessment would be fruitful.

Theoretical Framework 1: Dimensions of Systems Thinking

Grohs et al. (2018) describe a three-dimensional framework of systems thinking (Fig. 1). The *problem dimension* considers both technical elements and contexts in analyzing a complex problem, including assumptions, goals, and constraints. The *perspective dimension* considers multiple perspectives or "frames" of the problem across potential stakeholders. The *time dimension* considers the history of a situation and considers potential short- and long-term unexpected consequences of each possible action.

This framework was developed based on the literature of systems thinking research and is an example of a framework of systems thinking as a general perspective. A scenario-based assessment of systems thinking based on this framework was developed by Grohs et al. (2018) which presents participants with a short scenario and asks them to respond to six questions related to problem identification, information needs, stakeholder awareness, goals, unintended consequences, and implementation challenges. A rubric is used to

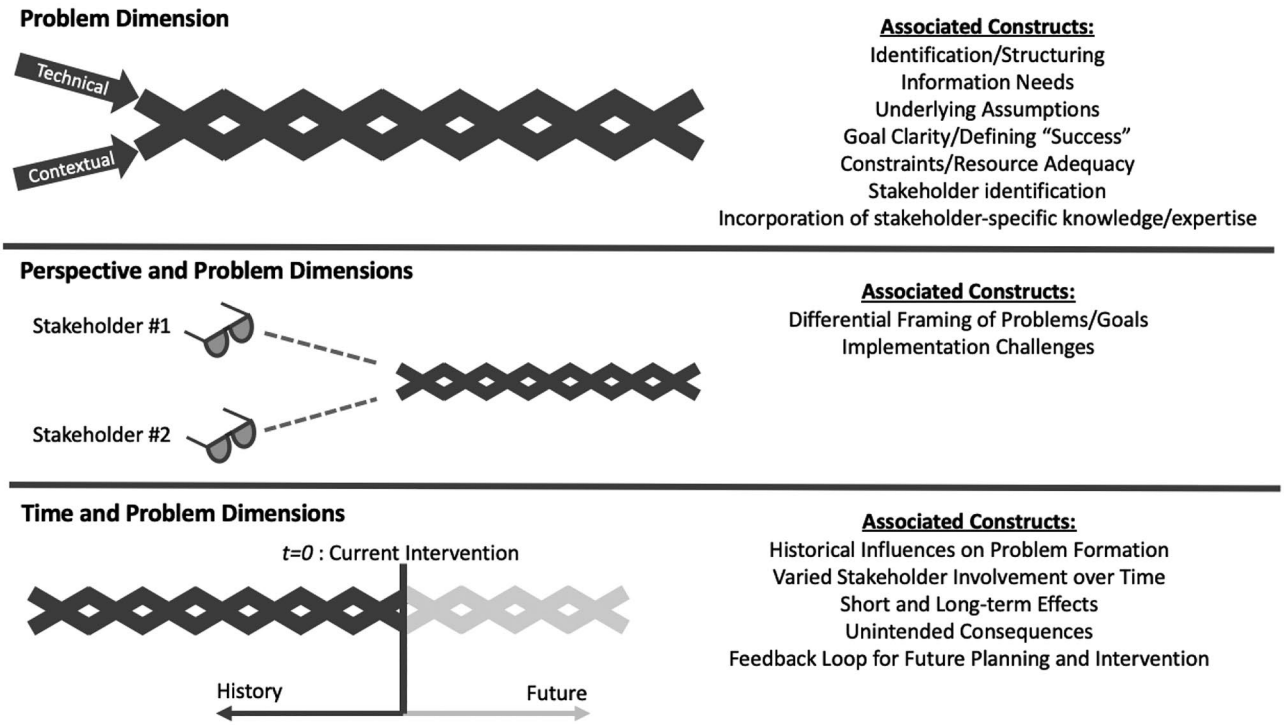

Fig. 1 Dimensions of Systems Thinking framework (reproduced from Grohs et al., 2018)

score these six responses individually and also for alignment across responses (Grohs et al., 2018).

Theoretical Framework 2: Systems as Webs of Interconnections

Another approach to understanding systems thinking is to consider the basic criteria that differentiate a systems approach from its alternatives. From this perspective, systems thinking focuses on understanding systems as a whole and recognizing interconnections between different system parts (Meadows, 2008; Senge, 2006). When thinking about systems this way, one can see that problems are often related and changing overtime (Ackoff, 1971, 1994; Senge, 2006) and uncover circular chains of causes and effects (*feedback loops*) within the system (Senge, 2006; Sterman, 2000). It is this foundational understanding that can help individuals move from a focus on individual events to changing patterns over time and from blaming individuals or external enemies to seeing ourselves as part of the system potentially contributing to the problem (Meadows, 2008; Senge, 2006). By seeing interconnections, people recognize that for every action, there will be a reaction from the system. In contrast, a non-systems thinker sees events linearly, X causing Y, and looks for easy and fast solutions to fix symptoms of problems (Forrester, 1971). Such simplistic solutions often fail due to the system's resistance and result in unintended consequences (Forrester, 1971; Senge, 2006). The recognition of interconnections and feedback structures is the foundation of many systems thinking schools of thought, such as system dynamics (Ghaffarzadegan & Larson, 2018; Randers, 2019; Richardson, 2011; Sterman, 2018).

Using this framework, we can compare and contrast individuals' mental models to understand whether they see a complex socio-environmental problem as a problem caused by a single factor/player or are able to recognize how the actions of different players are related. To that end, a scenario-based assessment based on a real-world complex case was designed and used to assess individuals' evaluation of the problem (Davis et al., 2020). Similar to the assessment described in the previous section, participants are presented with a scenario describing the situation and are asked to explain what went wrong. A scoring process is then used to identify the number of variables, causal links, and feedback loops described in the response.

Methods and Results

This paper describes data collected from two related studies, each built on one of the previously described theoretical frameworks of systems thinking. For both studies, participants were students enrolled in a spring semester first-year engineering course called *Global Engineering Practice*. Because the university has a common first-year engineering program, all participants were General Engineering majors at the time of data collection, typically in their first year in college following high school (with a small number of older transfer students); those students then joined different engineering departments the following year (e.g., mechanical engineering, civil engineering, etc.). The demographic composition of engineering students who self-select into this class tend to be more diverse than the College of Engineering, particularly with respect to gender, as about half of the class enrollment was women compared to about 20–25% of the college more broadly. Students from racially minoritized groups were slightly more represented in the class relative to the college. All students included in the sample consented to participate in this study, which has been approved by the university IRB office. We present the Methods and Results of Study 1 first followed by the Methods and Results of Study 2 since we followed a sequential approach to data collection and analysis.

Study 1 Data Collection and Analysis (Dimensions of Systems Thinking Framework)

The data for this study were taken from the 2017 ($n = 123$) and 2018 ($n = 140$) iterations of the course for a total of 263 participants. An instrument was administered to all students in class that included a scenario-based assessment of systems thinking aligned with the *Dimensions of Systems Thinking* framework. This assessment presents students with a one-paragraph description of a complex situation facing the Village of Abeesee. Students then complete six open-ended questions aligned with the dimensions of the framework. This scenario is scored using a rubric where students are rated between 0 (irrelevant response) and 3 (strong response) on each of the six questions and also on the extent to which their responses align logically across questions. All assessments were scored by a single researcher who had undergone training from the developers of the instrument. This researcher discussed multiple subsets of the scored data with the instrument developers as a formal peer audit process. Strong responses are characterized by a holistic framing of the problem, including both technical and contextual details, an acknowledgement of short-term and long-term considerations, and the inclusion of a variety of stakeholders. Intermediate responses include only some of these aspects and are typically limited in their analysis of the scenario. For a detailed description of the scenario, the rubric, and their development process, see Grohs et al. (2018). The text of the scenario is included in Appendix 1.

This scoring process yields seven variables for analysis purposes (i.e., problem identification, information needs,

stakeholder awareness, goals, unintended consequences, implementation challenges, and alignment; an overall score is not calculated). Students also completed four self-report assessments related to systems thinking (Lattuca et al., 2013; Moore et al., 2010; Ro et al., 2015; Sosu, 2013). The Systems Thinking Scale (Moore et al., 2010) was chosen because of its explicit focus on systems thinking. The other chosen assessments measure what we expect to be related constructs given the literature on complex ill-structured problem solving (described earlier in "Theoretical Perspectives Framing Systems Thinking Assessments" section) . Specifically, Jonassen (2010) highlights casual reasoning, analogical reasoning, and epistemological beliefs as cognitive skills that describe individual differences in ability to solve ill-structured problems. Thus, we would expect that there may be relationships between some of these assessments and behavioral measures of systems thinking ability. These instruments and their scales/subscales are shown in Table 1. The items for each scale are included in Appendix 2.

We conducted both correlation and regression analyses. First, we calculated a correlation matrix comparing the scenario assessment scores with the total scores and scale and/or sub-scale scores for each of the self-report assessments. Because we made multiple comparisons, we adjusted *p* values for family-wise error rate using the Holm correction, choosing this option because it provides a balance between reducing the risk of Type I errors and maintaining statistical power (Field et al., 2012). Although much of our data are Likert-scale survey responses, we used the Pearson correlation method as this has been shown to be robust for use with both ordinal and non-normal data (Norman, 2010). Next, we conducted multiple regression analyses with the scenario scores as the dependent variable and the systems thinking self-report scores as the independent variables. For each regression analysis, we checked for multicollinearity using the variance inflation factor (VIF) and for independent errors using the Durbin-Watson test. All VIF values were under 2 and Durbin-Watson values were between 1.75 and 2.25, both of which are well within the recommended values (Field et al., 2012).

Results for Study 1

The correlation matrix comparing the scores for the seven dimensions of the Dimensions of Systems Thinking scenario assessment with the scale and sub-scale scores for the four self-report assessments is shown in Fig. 2 (the table of *p* values for this matrix is in Appendix 3). There are few significant correlations between the scores on the different questions of the scenario assessment: a medium correlation between unintended consequences and implementation challenges and a weak correlation between goals and alignment (where weak correlation is $0.1 < r < 0.3$, medium is $0.3 < r < 0.5$, and strong is $r > 0.5$; Field et al., 2012). There is no specific expectation from the Grohs et al. (2018) instrument that strong correlations would be seen across these constructs in a mixed population of respondents. For an expert respondent, it would make sense that all constructs would be similarly high, but it is not clear that any prescribed patterns should exist for novice respondents. The relationship observed here between unintended consequences and implementation challenges could suggest a pattern worth investigating, but it could also be explained by those two constructs both being scored from respondent text to the same prompt. One takeaway from our analysis is that for undergraduate student respondents, we do not see strong correlations between constructs, and this observation may suggest that the constructs are indeed measuring different things. Even more distinct is the complete lack of significant correlations between the scenario assessment scores and the self-report scales while also observing strong correlations among the self-report scales themselves (particularly sub-scales within the Critical Thinking Disposition Scale). The Contextual Competence Scale was the most differentiated from the others, with only medium correlations

Table 1 Self-report assessments used for comparison in Study 1

Scale/sub-scale	# of items	Cronbach's alpha	Sample item
Systems Thinking Scale (Moore et al., 2010)			
N/A	20	0.89	I keep in mind that proposed changes can affect the whole system
Critical Thinking Disposition Scale (Sosu, 2013)			
Reflective skepticism	4	0.76	I often re-evaluate my experiences so that I can learn from them
Critical openness	7	0.76	I am often on the lookout for new ideas
Interdisciplinary Competence Scales (Lattuca et al., 2013)			
Interdisciplinary skills	8	0.86	I enjoy thinking about how different fields approach the same problem in different ways
Reflective behavior	2	0.74	I frequently stop to think about where I might be going wrong or right with a problem solution
Contextual Competence Scale (Ro et al., 2015)			
N/A	4	0.80	Please rate your ability to recognize how different contexts can change a solution

Fig. 2 Study 1 correlation matrix

Variable	*n*	*M*	*Min*	*Max*	*SD*	1A	1B	1C	1D	1E	1F	1G	2	3A	3B	3C	4A	4B	5
Scenario Scores:																			
1A - Problem Identification	263	1.62	0	3	0.64	–													
1B - Information Needs	263	1.81	0	3	0.53	0.12	–												
1C - Stakeholder Awareness	263	1.24	0	3	0.99	-0.03	-0.01	–											
1D - Goals	263	1.71	0	3	0.62	0.06	0.04	0.06	–										
1E - Unintended Consequences	263	1.38	0	3	0.58	0.01	0.09	0.16	0.16	–									
1F - Implementation Challenges	263	1.64	0	3	0.57	0.07	0.11	0.12	0.03	0.33***	–								
1G - Alignment	263	1.71	0	3	1.00	-0.09	0.09	0.01	0.25**	0.07	0.05	–							
Self-Report Scales:																			
2 - Systems Thinking Scale	258	81.56	33	100	9.15	-0.02	0.10	-0.05	-0.02	-0.03	0.07	0.02	–						
3A - Critical Openness Sub-Scale	257	30.08	21	35	4.16	0.03	0.05	-0.01	-0.03	0.03	0.08	0.01	0.51***	–					
3B - Reflective Skepticism Sub-Scale	257	16.76	10	20	2.81	0.03	-0.01	-0.04	0.04	0.01	0.05	0.04	0.51***	0.76***	–				
3C - Critical Thinking Disposition (3A+3B)	257	46.84	33	55	6.56	0.03	0.03	-0.02	0.00	0.02	0.07	0.02	0.54***	0.96***	0.91***	–			
4A - Interdisciplinary Skills Sub-Scale	257	4.29	2.4	5	0.56	0.04	0.06	0.00	-0.05	0.02	0.02	0.04	0.53***	0.52***	0.48***	0.53***	–		
4B - Reflective Behavior Sub-Scale	257	4.23	1	5	0.67	-0.09	-0.02	0.04	-0.02	0.00	0.08	0.03	0.51***	0.46***	0.56***	0.53***	0.52***	–	
5 - Contextual Competence Scale	257	3.54	1	5	0.68	0.01	0.09	0.04	0.05	-0.03	0.03	0.04	0.40***	0.32***	0.40***	0.37***	0.40***	0.37***	–

Note. *$p < .05$. **$p < .01$. ***$p < .001$. Holm correction used to adjust for multiple comparisons.

across the board. Overall, however, we observe that students' scores on the self-report scales align with each other but not with the Dimensions of Systems Thinking scenario assessment scores.

Multiple linear regression was conducted with each of the seven dimensions as the dependent variable and the scores for the four self-report assessments as independent variables. As shown in Table 2, only the regression for the problem identification dimension revealed any significant relationships (three of the dimensions are shown as examples given that the results were similar across dimensions. The results of the remaining dimensions are included in Appendix 4).

Our main finding from these analyses is that the overall models are not statistically significant for any of the seven dimensions and the adjusted *R*-squared values are all quite small, indicating that the models explain almost none of the variation in students' scores on the Dimensions of Systems Thinking assessment. Although one significant predictor was identified, the overall model performance clearly indicates that students' scores on the self-report assessments do not predict their scores on this scenario-based assessment.

Study 2 Data Collection and Analysis (Systems as a Web of Interconnections)

The data for this study were taken from the 2019 ($n = 155$) iteration of the course. This study followed a similar structure to Study 1 except that the scenario assessment used was based on the Systems as a Web of Interconnections framework described earlier. The Lake Urmia Vignette (LUV) provides a four-paragraph description of a lake that has dried up over time and related economic, environmental, social, and political events and outcomes connected to the lake (Davis et al., 2020). Students are asked to respond to one question asking them to "Describe the problems facing Lake Urmia in detail and explain why the lake shrank over the years." Most students write about a paragraph in response to this question, and these responses are analyzed to identify constructs related to the Systems as a Web of Interconnections framework. First, students receive points for each unique variable they identify as part of the Lake Urmia system (e.g., local population). Second, they receive points for connecting these variables together through causal links (e.g., the population uses lake water for irrigation). Finally, students receive points for identifying feedback loops where the causal relationships connect to each other (e.g., the population uses lake water for irrigation, which increases the available food, resulting in an increase in the population). Each students' points are totaled to calculate their overall score on the scenario. In our scoring process, each response was scored by two independent raters who then compared their results and discussed until scores were agreed upon. For more information about the development

Table 2 Regression results for three of the dimensions of systems thinking

Variable	Problem identification		Information needs		Stakeholder awareness	
	Beta	*p* value	Beta	*p* value	Beta	*p* value
(Intercept)		< 0.001***		< 0.001***		0.012*
STS Score	− 0.04	0.633	0.12	0.143	− 0.12	0.163
CTD Score	0.08	0.321	− 0.02	0.768	− 0.03	0.686
ID Score	0.09	0.265	0.04	0.659	0.01	0.867
RB Score	− 0.17	0.039*	− 0.11	0.160	0.09	0.276
CC Score	0.02	0.759	0.08	0.267	0.06	0.414
Adj. *R*-squared	0.002		0.003		− 0.006	
p value	0.359		0.330		0.647	

STS Systems Thinking Scale, *CTD* critical thinking disposition, *ID* interdisciplinary skills, *RB* reflective behavior, *CC* contextual competence

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

and scoring of the LUV scenario, see Davis et al. (2020). The text of the scenario is included in Appendix 5.

In addition to the LUV scenario, one of the systems thinking self-report assessments and a social desirability scale (Steenkamp et al., 2010) were included in the survey used in Study 2. These instruments and their subscales are shown in Table 3. The items for each scale are included in Appendix 6.

These changes to the self-report assessments included from Study 1 to Study 2 were informed by the results of Study 1. Because the various self-report assessments were strongly correlated with each other in Study 1, we decided only one was needed in Study 2. The results of Study 1 also suggested that perhaps a construct like social desirability could influence students' responses on the self-report assessments; that is, students may choose answers based on what they wish to be true rather than what they believe to be true. To explore this possibility, we included the Balanced Inventory of Desirable Responding (BIDR) in Study 2. If scores on the BIDR significantly correlated with the self-report assessment scores, this relationship would suggest that there may be a social desirability bias in students' responses.

Lastly, to explore other possible skills that might relate to strong responses on the scenario-based assessment, we collected a few more variables. First, we asked students to self-rate their math ability relative to the average engineering student because both math and systems thinking involve complex thinking. Second, we had students respond to a basic question about feedback loops to determine their familiarity with concepts from the Systems as a Web of Interconnections framework. Third, we counted the number of words in students' responses, as it is possible that students who wrote more would achieve better scores based on the scoring system used in this study.

We followed similar approaches as in Study 1 to analyze data. In addition, in Study 2, we conducted the same initial regression analysis and then two additional analyses: (1) adding the social desirability scales and (2) adding the background knowledge and word count variables as independent variables. As discussed previously, we used the Holm correction in the correlation analysis to adjust for multiple comparisons (Field et al., 2012). For each regression analysis, we checked for multicollinearity using the variance inflation factor (VIF) and for independent errors using the Durbin-Watson test. All VIF values were under 2 and Durbin-Watson values were between 1.75 and 2.25, both which are well within the recommended values (Field et al., 2012). We also included demographic variables in the regression models in an attempt to account for known differences in engineering student responses to these kinds of measures as demonstrated in prior research (e.g., Knight, 2014). We used demographic data that were collected by the institution, which at the time of data collection used male or female

Table 3 Self-report assessments used for comparison in Study 2

Sub-scale	# of items	Cronbach's alpha	Sample item
Critical Thinking Disposition Scale (Sosu, 2013)			
Reflective skepticism	4	0.62	I often re-evaluate my experiences so that I can learn from them
Critical openness	7	0.61	I am often on the lookout for new ideas
Balanced Inventory of Desirable Responding (Steenkamp et al., 2010)			
Egoistic response tendencies	10	0.58	I never regret my decisions
Moralistic response tendencies	10	0.65	I never cover up my mistakes

gender categories. We support changes in practices to how this demographic information is collected in the future to recognize that gender is a non-binary social construct.

Results for Study 2

The results from Study 1 led to our decision to only include one self-report assessment in Study 2 and include the social desirability scale to explore one possible explanation for the lack of alignment between the Dimensions of Systems Thinking scenario and self-report results. Another possible explanation is that the scenario and self-report assessments are not assessing the same or hypothesized related constructs. We therefore attempted to use a different theoretical framework and scenario assessment (the LUV scenario) in Study 2 to see if we would observe a different result.

The correlation matrix showing the relationships between the LUV scenario scores, a self-report assessment, a social desirability scale, and the background knowledge and word count variables is shown in Fig. 3 (the table of critical *p* values for this matrix can be found in Appendix 7). There are few significant correlations between the LUV scenario scores and the other variables. The only variable that is significantly related to the LUV scores is the word count of the students' responses. Students who wrote longer responses also identified more variables and more causal links (although word count is not correlated with the number of loops they identified). There are also significant correlations between the sub-scores for the LUV scenario, but once again, loops are less strongly related to the other sub-scores. Although the critical thinking disposition scores do not relate to the LUV scores, it is notable that they are somewhat related to students' self-rated math competence. There are also significant correlations between the subscales for both critical thinking disposition and social desirability.

Multiple linear regression analyses were conducted with each of the LUV scenario sub-scores and total score as the dependent variables. Three regressions were run for each of the four dependent variables: (1) demographic variables (age and gender); (2) adding the self-report and social desirability instruments; and (3) adding the background knowledge and word count variables. The background knowledge variables were the only significant predictors across all analyses, although this finding varied somewhat between the LUV sub-scores. Table 4 shows the results of these analyses for the LUV total score variable.

Because the results were similar for the sub-scores (variable, causal link, and causal loop identification), we do not show those results here (see Appendix 8). The most notable difference in results across the dependent variables was that the regression 3 model was only

Fig. 3 Study 2 correlation matrix

Variable	*n*	*M*	*Min*	*Max*	*SD*	1A	1B	1C	1D	2A	2B	2C	3A	3B	4	5	6
Scenario Scores:																	
1A - Number of Variables Identified	143	10.59	4	23	4.00	–											
1B - Number of Causal Links Identified	143	9.17	3	25	3.97	0.94***	–										
1C - Number of Loops Identified	143	0.16	0	4	0.45	0.21	0.29*	–									
1D - Total Scenario Score	143	20.08	7	51	8.13	0.98***	0.98***	0.35***	–								
Systems Thinking Self-Report:																	
2A - Reflective Skepticism Sub-Scale	143	17.12	10	20	2.00	0.13	0.11	0.06	0.13	–							
2B- Critical Openness Sub-Scale	143	30.86	23	35	2.41	0.04	0.05	0.12	0.06	0.47***	–						
2C - Critical Thinking Disposition (2A+2B)	143	47.98	36	55	3.78	0.09	0.09	0.11	0.10	0.83***	0.88***	–					
Inventory of Desirable Responding:																	
3A - Egoistic Response Tendencies	143	29.81	18	40	4.55	-0.14	-0.07	0.05	-0.09	0.18	0.12	0.17	–				
3B - Moralistic Response Tendencies	143	32.71	18	49	5.35	0.15	0.16	0.18	0.18	0.18	0.16	0.20	0.31**	–			
Comparison Variables:																	
4 - Score on Basic Feedback Problem	143	0.98	0	3	0.72	0.18	0.22	0.16	0.21	-0.05	0.01	-0.02	-0.01	0.15	–		
5 - Self-Rated Math Competence	142	3.75	2	5	0.74	0.10	0.04	0.00	0.07	0.28*	0.17	0.26	0.06	0.08	-0.02	–	
6 - Word Count for Response	143	108.5	31	240	41.43	0.65***	0.66***	0.30*	0.68***	0.20	0.16	0.20	-0.09	0.12	0.04	-0.03	–

Note. *$p < .05$. **$p < .01$. ***$p < .001$. Holm correction used to adjust for multiple comparisons.

Table 4 Regression results for LUV total score

	Regression 1		Regression 2		Regression 3	
Variable	Beta	*p* value	Beta	*p* value	Beta	*p* value
(Intercept)		0.450		0.642		0.711
Age	0.01	0.930	−0.01	0.916	−0.01	0.890
Gender (1 = woman)	0.06	0.488	0.00	0.965	0.03	0.647
CTD score			0.09	0.281	−0.07	0.308
ERT score			−0.18	0.052	−0.05	0.453
MRT score			0.21	0.019*	0.08	0.211
Feedback score					0.17	0.005**
Self-rate math					0.11	0.092
Word count					0.67	< 0.001***
Adj. *R*-squared	−0.011		0.030		0.482	
p value	0.779		0.104		< 0.001***	

CTD critical thinking disposition, *ERT* egoistic response tendencies, *MRT* moralistic response tendencies
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

significant at the $p < 0.05$ level for the causal loop variable, and the adjusted *R*-squared was negligible.

The regression analyses revealed that word count was a primary factor in students' scores, but even after accounting for this relationship, students' scores on a basic feedback problem were still a significant predictor of their total LUV score. In considering multiple regression of the sub-scores of the LUV assessment shown in Appendix 8, students' self-rated math competence barely meets the $p < 0.05$ criteria as an additional significant predictor for the number of variable sub-score and students' scores on the feedback problem related to both their identification of variables and causal links (but not loops). It is promising that the feedback problem score is a relevant variable because it suggests that the LUV scenario is capturing some understanding of concepts related to the Systems as a Web of Interconnections framework. However, it remains unclear what factors may influence students' ability to identify loops within the LUV scenario.

Discussion

Our study explored the relationship between students' scores on self-report assessments of constructs expected to be related to systems thinking and scenario-based assessments of systems thinking ability. Following Pike's (2011) suggestion to use theory to inform analysis of self-report assessments, we used scenario-based assessments that were developed based on two different theoretical frameworks of systems thinking. Through two sequential studies following the same research approach (with different scenarios), our results revealed no significant relationships between students' performance on these scenarios and their scores on the self-report assessments. These findings remained consistent in both correlation and regression analyses. In Study 2, we found that students' performance on a feedback loop problem and word count of their scenario response significantly related to their scenario assessment scores. These variables accounted for a large portion of the variation in the scenario assessment scores, whereas the self-report assessment scores were insignificant. These results could indicate that the scenarios assess a different construct than the self-report assessments, or that they are assessing the same construct at a different level of granularity. In either case, these two forms of assessment do not appear to be in alignment with each other despite their theoretical linkages related to students' systems thinking ability.

This study contributes to the ongoing discussion about the effectiveness of self-report measures in educational research. We build on prior work suggesting that self-report measures may be reasonable in some contexts and for some constructs but not for others (Chan, 2009). Previous discussions have focused on self-reporting learning gains and attitudes, suggesting that the former is not effective, whereas the latter may be best assessed through self-reports (Chan, 2009; Porter, 2011). In this study, we explored competence as another type of learning outcome that is often assessed using self-report assessments, but which has been explored less thoroughly in the literature. Such instruments are common in educational research beyond the systems thinking and problem-solving space that we focused on in this study. For example, outcomes like intercultural competence (e.g., Braskamp et al., 2014; Hammer et al., 2003), leadership (e.g., Novoselich & Knight, 2017), and civic attitudes and skills associated with community engagement (e.g., Kirk & Grohs, 2016; Moely et al., 2002; Reeb et al., 2010) are also frequently assessed using this approach. Other authors have pointed out the lack of evidence for construct validity for many of these instruments (Lattuca et al., 2013). Some prior work has revealed

that students with more experience with the competencies in question actually decline on the self-report assessments, suggesting that as they become more familiar with the subject, students realize how little they actually know. This aligns with the more general findings that experts can more accurately assess their competence than novices (Ehrlinger et al., 2008; Kruger & Dunning, 1999). Prior studies of intercultural competence have revealed similar results to the current study, for example, comparing self-report scores to both scenario-based assessments and qualitative analysis of student journals and finding that the self-report scores did not correlate with these more direct forms of assessment (Davis, 2020). One study even found significant negative correlations between scores on a scenario-based assessment and scores on self-report assessments for practicing engineers (Jesiek et al., 2020). In conjunction with this prior work, our study has the potential to inform the use of self-report assessments for both assessment and research purposes.

Limitations

One limitation of this study is that we do not have data comparing students' performance on the two scenario-based assessments, so we can make no claims about whether these two instruments are assessing the same aspects of systems thinking. Other recent research has begun to make such comparisons (e.g., Joshi et al., 2022), but more work is needed in this direction. A related second limitation is that with the exception of one of the assessments, most of the self-report assessments used in this study do not purport to measure systems thinking but rather related constructs. Thus, an alternate explanation for the lack of observed relationships between the self-report assessments and the scenario-based assessments could be that that the scenario-based measures do not have enough validity evidence beyond their original publications or that systems thinking ability does not have a relationship with ill-structured problem solving cognitive skills as hypothesized by Jonassen (2010). Further, our study used selected self-report assessments, but countless other tools could have been compared. The recent work of Dugan et al. (2022) systematically identifying a range of systems thinking assessment tools can inform future investigations exploring relationships between tools aiming to measure the same or similar constructs.

A third limitation is the sample used in our studies, which includes only first-year engineering students. As discussed in the literature review, more novice systems thinkers may be more inclined to overestimate their abilities on self-report assessments, so our data from first-year students could include greater inflation of their abilities than if we included more advanced students. On the other hand, our student sample may not differ much from a sample of other students when compared to much more advanced systems thinkers such as, for example, professional engineers (Mazzurco & Daniel, 2020; Mosyjowski et al., 2021). Our sample for both studies was also more diverse than the college of engineering in which it was situated and engineering programs broadly, especially in terms of gender. We have no reason to believe that systems thinking competence is related to gender, and gender was not a significant predictor in any of our regression models in Study 2. Nevertheless, it may be important to consider this aspect of our sample when comparing our findings to other contexts. Finally, our sample includes only engineering students, which represents only a small subset of students with whom these assessments can be used. Although we have no indication that engineering students would be better or worse at self-reporting their own abilities than other students, future research should expand on this work to explore whether there are differences across disciplines.

Broader Implications

This study suggests that further research is needed to understand self-report assessments of competence. In this study, we compared the self-report assessments to scenario-based assessments, but more expansive assessments could be pursued to further explore these results, such as having students complete a more in-depth systems thinking activity or project. A second need is for similar studies to be conducted with other self-report assessments of competence, such as for intercultural competence, critical thinking, or creative thinking. Anderson et al. (2017) provide one example of a study that reveals weaknesses with self-report assessments for both global citizenship and creative thinking. However, further research is needed to support both the claims of biases in self-report assessments and determine if these biases are constant or variable across self-report assessments for different competencies. Third, researchers should pursue the development and validation of alternative assessment approaches (e.g., scenario-based assessments, situational judgment tests) for competencies such as systems thinking. This paper builds on the development of two scenario-based assessments that require further validation through broader use and research. Such assessments have the benefit that they can be used as an instructional tool in addition to an assessment method, providing additional benefits that are lacking with self-report assessments.

Beyond assessment, however, our study and the other literature exploring this topic suggest that perhaps educators' attempts at assessment could be informed by a better understanding of competence. One understanding of competence presented by Lucia and Lepsinger (1999) suggests that it is made up of a combination of inherent aptitudes and characteristics, learnable skills and knowledge, and manifested

behaviors. Although self-report assessments could be used to assess some aspects of this definition (e.g., knowledge), they do not provide the ability to assess others (e.g., behaviors). This framework for understanding competence also suggests that certain aspects of competence may be context-specific, whereas others are applicable across contexts (Lucia & Lepsinger, 1999). Future research that explores the nature of competence and competence-development may also be needed before we can understand what we are assessing using different approaches and ensure that we are interpreting our various assessments accurately (Figs. 4 and 5, Tables 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14).

Conclusion

This study explored the relationships between students' performance on self-report assessments and scenario-based assessments of systems thinking, finding that there were no significant relationships between the two assessment techniques. These results call into question the extensive use of self-report assessments as a method to assess systems thinking and other related competencies in educational research and evaluation. Future work should explore these findings further and support the development of alternative formats for assessing competence.

Appendix 1. The Village of Abeesee scenario

The Village of Abeesee has about 50,000 people. Its harsh winters and remote location make heating a living space very expensive. The rising price of fossil fuels has been reflected in the heating expenses of Abeesee residents. Many residents are unable to afford heat for the entire winter (5 months). A University of Abeesee study shows that 38% of village residents have gone without heat for at least 30 winter days in the last 24 months. Last year, 27 Abeesee deaths were attributed to unheated homes. Most died from hypothermia/exposure (21), and the remainder died in fires or from carbon monoxide poisoning that resulted from improper use of alternative heat sources (e.g., burning trash in unventilated space).

Appendix 2. Items for Scales Used in Study 1

Table 5 Systems Thinking Scale (Moore et al., 2010)

Item #	Item
STS1	I seek everyone's view of the situation
STS2	I look beyond a specific event to determine the cause of the problem
STS3	I think understanding how the chain of events occur is crucial
STS4	I include people in my work unit to find a solution
STS5	I think recurring patterns are more important than any one specific event
STS6	I think of the problem at hand as a series of connected issues
STS7	I consider the cause and effect that is occurring in a situation
STS8	I consider the relationships among coworkers in the work unit
STS9	I think systems are constantly changing
STS10	I propose solutions that affect the work environment, not specific individuals
STS11	I keep in mind that proposed changes can affect the whole system
STS12	I think more than one or two people are needed to have success
STS13	I keep the mission and purpose of the team/organization in mind
STS14	I think small changes can produce important results
STS15	I consider how multiple changes affect each other
STS16	I think about how different team/organization members might be affected by the improvement
STS17	I try strategies that do not rely on people's memory
STS18	I recognize system problems are influenced by past events
STS19	I consider the past history and culture of the team/organization unit
STS20	I consider that the same action can have different effects over time, depending on the state of the system

Table 6 Critical Thinking Disposition Scales (Sosu, 2013)

Item #	Item
Reflective Skepticism Sub-Scale (RS)	
RS1	I usually think about the wider implications of a decision before taking action
RS2	I usually check the credibility of the source of information before making judgements
RS3	I often re-evaluate my experiences so that I can learn from them
RS4	I often think about my actions to see whether I could improve them
Critical Openness Sub-Scale (CO)	
CO1	I am often on the lookout for new ideas
CO2	I often use new ideas to shape (modify) the way I do things
CO3	I use more than one source to find out information for myself
CO4	It is important to justify the choices I make
CO5	It's important to understand other people's viewpoint on an issue
CO6	I usually try to think about the bigger picture during a discussion
CO7	I sometimes find a good argument that challenges some of my firmly held beliefs

Table 7 Contextual Competence Scale (Ro et al., 2015)

Item #	Item
CC1	Ability to use what you know about different cultures, social values, or political systems in engineering solutions
CC2	Ability to recognize how different contexts can change a solution
CC3	Knowledge of contexts that might affect the solution to an engineering problem
CC4	Knowledge of the connections between technological solutions and their implications for whom it benefits

Table 8 Interdisciplinary Competence Instrument (Lattuca et al., 2013)

Item #	Item
Interdisciplinary Skills Scale (ID)	
ID1	I value reading about topics outside of engineering
ID2	I enjoy thinking about how different fields approach the same problem in different ways
ID3	Not all engineering problems have purely technical solutions
ID4	In solving engineering problems I often seek information from experts in other academic fields
ID5	Given knowledge and ideas from different fields, I can figure out what is appropriate for solving a problem
ID6	I see connections between ideas in engineering and ideas in the humanities and social sciences
ID7	I can take ideas from outside engineering and synthesize them in ways to better understand a problem
ID8	I can use what I have learned in one field in another setting or to solve a new problem
Reflective Behavior Scale (RB)	
RB1	I often step back and reflect on what I am thinking to determine whether I might be missing something
RB2	I frequently stop to think about where I might be going wrong or right with a problem solution

Appendix 3. Study 1 *p*-values for Correlation Matrix

Fig. 4 Study 1 *p* values for correlation matrix

Variable	1A	1B	1C	1D	1E	1F	1G	2	3A	3B	3C	4A	4B	5
Scenario Scores:														
1A - Problem Identification	–													
1B - Information Needs	1	–												
1C - Stakeholder Awareness	1	1	–											
1D - Goals	1	1	1	–										
1E - Unintended Consequences	1	1	.584	.624	–									
1F - Implementation Challenges	1	1	1	1	<.001	–								
1G - Alignment	1	1	1	.003	1	1	–							
Self-Report Scales:														
2 - Systems Thinking Scale	1	1	1	1	1	1	1	–						
3A - Critical Openness Sub-Scale	1	1	1	1	1	1	1	<.001	–					
3B - Reflective Skepticism Sub-Scale	1	1	1	1	1	1	1	<.001	<.001	–				
3C - Critical Thinking Disposition (3A+3B)	1	1	1	1	1	1	1	<.001	<.001	<.001	–			
4A - Interdisciplinary Skills Sub-Scale	1	1	1	1	1	1	1	<.001	<.001	<.001	<.001	–		
4B - Reflective Behavior Sub-Scale	1	1	1	1	1	1	1	<.001	<.001	<.001	<.001	<.001	–	
5 - Contextual Competence Scale	1	1	1	1	1	1	1	<.001	<.001	<.001	<.001	<.001	<.001	–

Note. $*p < .05$. $**p < .01$. $***p < .001$. Holm correction used to adjust for multiple comparisons.

Appendix 4. Additional regression results from Study 1

Table 9 Regression results for two of the dimensions of systems thinking

	Goals		Unintended consequences	
Variable	Beta	*p* value	Beta	*p* value
(Intercept)		< .001***		< .001***
STS score	− 0.02	.856	− 0.06	.474
CTD score	0.03	.732	0.05	.554
ID score	− 0.08	.327	0.05	.514
RB score	− 0.01	.860	− 0.01	.875
CC score	0.08	.270	− 0.04	.535
Adj. *R*-squared	− 0.012		− 0.014	
p value	.836		.922	

STS Systems Thinking Scale, *CTD* critical thinking disposition, *ID* interdisciplinary skills, *RB* reflective behavior, *CC* contextual competence

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 10 Regression results for two of the dimensions of systems thinking

	Implementation challenges		Alignment	
Variable	Beta	*p* value	Beta	*p* value
(Intercept)		< .001***		.030*
STS score	0.04	.658	0.00	.956
CTD score	0.05	.579	− 0.01	.919
ID score	− 0.06	.492	0.03	.723
RB score	0.07	.414	0.01	.923
CC score	− 0.01	.936	0.03	.631
Adj. *R*-squared	− 0.010		− 0.017	
p value	.793		.985	

STS Systems Thinking Scale, *CTD* critical thinking disposition, *ID* interdisciplinary skills, *RB* reflective behavior, *CC* contextual competence

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Appendix 5. Lake Urmia Vignette

Lake Urmia in north-west Iran was the largest lake in the Middle East and the sixth largest saltwater lake on earth. Unique ecological features of this UNESCO Biosphere Reserve had made Urmia the largest habitat of brine shrimp. The lake is the home of various species of Archaebacteria and bacteria, microfungi, phytoplankton, and 311 species of plants. It also hosts 226 kinds of birds. The lake is also known for its hundreds of small islands serving as stopover points during the migrations of several wild birds to and from Russia. The town Urmia, on the west side of the lake, has a population of 700,000 people.

Several reports show that the lake suffers from serious ecological problems, and many of the indictors are easily observable from the lake itself. Between 1972 and 2014, the area of the lake shrank by 88%. The evaporation of the water has exposed the lakebed and caused windblown salt, which may lead to environmental health crises, including increase in infant mortality, cancer, and liver, kidney, and respiratory diseases. This phenomenon is similar to what happened after the death of the Aral Sea. In addition, it will increase unemployment by reducing tourism and shrinking the fertility of the land in the region.

Fortunately, the public awareness about the lake has increased and huge outcries urged the government to take action. Government officials promised to spend $5 million to save the lake in a period of 10 years. Officials attribute the lake's desiccation to the drought in recent years while critics point to mismanagement of water resources and construction of a raised road across the lake. The population of the region more than tripled during the past 40 years. Multiple dam construction and pipe transfer projects have made water available for domestic and agricultural purposes. In 1999, a project was completed to pump water from the Zarineh River (one of the main feeders of the lake) to the largest city of the area. In addition, forests have been transferred to agricultural lands to fulfill the needs of the growing population. Specifically, the forest cover of Zagros Mountain has declined. Forests of Zagros maintain naturally controlled water flow to rivers feeding the lake. As the lake shrinks, the climate of the region becomes drier and more water is needed for agriculture.

In 2011, multiple demonstrations took place in cities close to the lake demanding that the government take immediate actions to protect Urmia Lake. Not all demonstration ended peacefully. The slogan "let me cry to fill the lake," a highly chanted motto, depicted the emotional reaction of the region. According to official state reports, at least 70 supporters of the lake were arrested. Several proposals were discussed including channeling water from other rivers to the lake, destroying several dams, or funding relocation of people living around the lake. In order to find solutions for saving the lake, policy makers need to know what caused the lake to shrink by 88% in 44 years. The lake is just an example of many similar environmental challenges that humans are dealing with especially in less developed regions.

Appendix 6. Items for Scales Used in Study 2

Table 11 Balanced Inventory of Desirable Responding (Steenkamp et al., 2010)

Item #	Item
Egoistic response tendencies (ERT)	
ERT1	My first impressions of people usually turn out to be right
ERT2	It would be hard for me to break any of my bad habits (r)
ERT3	I have not always been honest with myself (r)
ERT4	I always know why I like things
ERT5	Once I've made up my mind, other people can seldom change my opinion
ERT6	It's hard for me to shut off a disturbing thought (r)
ERT7	I never regret my decisions
ERT8	I rarely appreciate criticism (r)
ERT9	I am very confident of my judgments
ERT10	I don't always know the reasons why I do the things I do (r)
Moralistic response tendencies (MRT)	
MRT1	I sometimes tell lies if I have to (r)
MRT2	I never cover up my mistakes
MRT3	I always obey laws, even if I am unlikely to get caught
MRT4	I have said something bad about a friend behind his or her back (r)
MRT5	When I hear people talking privately, I avoid listening
MRT6	I have received too much change from a salesperson without telling him or her (r)
MRT7	When I was young I sometimes stole things (r)
MRT8	I have done things that I don't tell other people about (r)
MRT9	I never take things that don't belong to me
MRT10	I don't gossip about other people's business

Items indicated with (r) are reverse-scored

Appendix 7. Study 2 *p*-values for Correlation Matrix

Fig. 5 Study 2 *p* values for correlation matrix

Variable	1A	1B	1C	1D	2A	2B	2C	3A	3B	4	5	6
Scenario Scores:												
1A - Number of Variables Identified	–											
1B - Number of Causal Links Identified	<.001	–										
1C - Number of Loops Identified	.685	.027	–									
1D - Total Scenario Score	<.001	<.001	<.001	–								
Systems Thinking Self-Report:												
2A - Reflective Skepticism Sub-Scale	1	1	1	1	–							
2B- Critical Openness Sub-Scale	1	1	1	1	<.001	–						
2C - Critical Thinking Disposition (2A+2B)	1	1	1	1	<.001	<.001	–					
Inventory of Desirable Responding:												
3A - Egoistic Response Tendencies	1	1	1	1	1	1	1	–				
3B - Moralistic Response Tendencies	1	1	1	1	1	1	.891	.008	–			
Comparison Variables:												
4 - Score on Basic Feedback Problem	1	.436	1	.534	1	1	1	1	1	–		
5 - Self-Rated Math Competence	1	1	1	1	.034	1	.112	1	1	1	–	
6 - Word Count for Response	<.001	<.001	.013	<.001	.779	1	.685	1	1	1	1	–

Note. $*p < .05$. $**p < .01$. $***p < .001$. Holm correction used to adjust for multiple comparisons.

Appendix 8. Additional regression results from Study 2

Table 12 Regression results for LUV Number of Variables

	Regression 1		Regression 2		Regression 3	
Variable	Beta	*p* value	Beta	*p* value	Beta	*p* value
(Intercept)		.442		.519		.580
Age	0.01	.891	− 0.01	.906	− 0.01	.910
Gender (1 = woman)	0.05	.574	− 0.01	.868	0.02	.808
CTD score			0.09	.288	− 0.07	.290
ERT score			− 0.22	.016*	− 0.10	.160
MRT score			0.21	.022*	0.09	.211
Feedback score					0.14	.028*
Self-rate math					0.14	.042*
Word count					0.64	< .001***
Adj. *R*-squared	− 0.012		0.036		0.446	
p value	.853		.076		< .001***	

CTD critical thinking disposition, *ERT* egoistic response tendencies, *MRT* moralistic response tendencies
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 13 Regression results for LUV Number of Causal Links

	Regression 1		Regression 2		Regression 3	
Variable	Beta	*p* value	Beta	*p* value	Beta	*p* value
(Intercept)		.473		.691		.754
Age	0.01	.946	− 0.01	.940	− 0.01	.903
Gender (1 = woman)	0.07	.409	0.03	.774	0.05	.475
CTD score			0.08	.354	− 0.07	.292
ERT score			− 0.14	.133	− 0.01	.844
MRT score			0.19	.040*	0.06	.394
Feedback score					0.18	.004**
Self-rate math					0.09	.181
Word count					0.66	< .001***
Adj. *R*-squared	− 0.009		0.014		0.454	
p value	.698		.223		< .001***	

CTD critical thinking disposition, *ERT* egoistic response tendencies, *MRT* moralistic response tendencies
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 14 Regression results for LUV Number of Causal Loops

	Regression 1		Regression 2		Regression 3	
Variable	Beta	*p* value	Beta	*p* value	Beta	*p* value
(Intercept)		.813		.677		.646
Age	−0.01	.907	−0.01	.928	−0.01	.893
Gender (1 = woman)	0.01	.898	−0.01	.891	−0.01	.872
CTD score			0.08	.370	0.03	.752
ERT score			−0.02	.859	0.04	.662
MRT score			0.17	.056	0.11	.207
Feedback score					0.14	.100
Self-rate math					−0.01	.871
Word count					0.28	.001**
Adj. *R*-squared	−0.014		0.004		0.080	
p value	.979		.352		.014*	

CTD critical thinking disposition, *ERT* egoistic response tendencies, *MRT* moralistic response tendencies
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Funding This work was supported by the National Science Foundation (Award #: EEC-1824594).

Data Availability The data sets supporting the current study are not publicly available because they are an excerpt of research in progress and because of our IRB agreements but are available from the corresponding author on reasonable request.

Declarations

Ethics Approval This study was reviewed and approved by the Institutional Review Board (IRB) at Virginia Polytechnic Institute and State University (Project #11–098).

Consent Statement Informed consent was obtained from all individual participants included in the study. The consent forms used in this study explicitly stated that the data collected would be used in research publications.

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ackoff, R. L. (1971). Towards a system of systems concepts. *Management Science, 17*(11), 661–671. https://doi.org/10.1287/mnsc.17.11.661

Ackoff, R. L. (1994). Systems thinking and thinking systems. *System Dynamics Review, 10*(2–3), 175–188.

ACPA College Student Educators International. (2008). *Toward a sustainable future: The role of student affairs in creating healthy environments, social justice, and strong economies*. American College Personnel Association.

Anderson, R., Thier, M., & Pitts, C. (2017). Interpersonal and intrapersonal skill assessment alternatives: Self-reports, situational-judgment tests, and discrete-choice experiments. *Learning and Individual Differences, 53*, 47–60. https://doi.org/10.1016/j.lindif.2016.10.017

Bowman, N. (2010). Assessing learning and development among diverse college students. *New Directions for Institutional Research, 145*, 53–71. https://doi.org/10.1002/ir.322

Bowman, N. (2011). Examining systematic errors in predictors of college student self-reported gains. *New Directions for Institutional Research, 150*, 7–19. https://doi.org/10.1002/ir.386

Bowman, N., & Hill, P. (2011). Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains. *New Directions for Institutional Research, 150*, 73–85. https://doi.org/10.1002/ir.390

Braskamp, L. A., Braskamp, D. C., Merrill, K. C., & Engberg, M. E. (2014). *Global perspective inventory (GPI): Its purpose, construction, potential uses, and psychometric characteristics*. Global Perspective Institute, Inc.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*(1), 60–77. https://doi.org/10.1037/0022-3514.90.1.60

Chan, D. (2009). So why ask me? Are self-report data really that bad? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences* (pp. 309–336). Routledge.

Connell, K. Y. H., Remington, S. M., & Armstrong, C. M. (2012). Assessing systems thinking skills in two undergraduate sustainability courses: A comparison of teaching strategies. *Journal of Sustainability Education, 3*, 1–15.

Dale, A., & Newman, L. (2005). Sustainable development, education and literacy. *International Journal of Sustainability in Higher Education, 6*(4), 351–362.

Davis, K. A. (2020). *Pursuing intentional design of global engineering programs: Understanding student experiences and learning outcomes* [Dissertation, Virginia Tech]. Retrieved February 17, 2023, from https://vtechworks.lib.vt.edu/handle/10919/97979

Davis, K. A., Ghaffarzadegan, N., Grohs, J. R., Grote, D., Hosseinichimeh, N., Knight, D. B., 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

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.

Dunning, D., & Kruger, J. (2002). Unskilled and unaware—But why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology, 82*(2), 189–192. https://doi.org/10.1037//0022-3514.82.2.189

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage Publications.

Forrester, J. W. (1971). Counterintuitive behavior of social systems. *Technological Forecasting and Social Change, 3*, 1–22. https://doi.org/10.1016/S0040-1625(71)80001-X

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

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

Hammer, M. R., Bennett, M. J., & Wiseman, R. (2003). Measuring intercultural sensitivity: The intercultural development inventory. *International Journal of Intercultural Relations, 27*, 421–443. https://doi.org/10.1016/S0147-1767(03)00032-4

Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*, 265–285. https://doi.org/10.1007/s11409-018-9185-6

Hu, M., & Shealy, T. (2018). Methods for measuring systems thinking: Differences between student self-assessment, concept map scores, and cortical activation during tasks about sustainability. Paper presented at 2018 ASEE Annual Conference & Exposition , Salt Lake City, Utah. https://doi.org/10.18260/1-2--30807

Jesiek, B. K., Woo, S. E., Parrigon, S., & Porter, C. (2020). Development of a situational judgement test (SJT) for global engineering competency (GEC). *Journal of Engineering Education*, 1–21. https://doi.org/10.1002/jee.20325

Jonassen, D. H. (2010). *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge.

Joshi, S. S., Davis, K. A., Czerwionka, L., Camps Troncoso, E., & Montalvo, F. J. (2022, June). A comparison of two scenario-based assessments of systems thinking. In *2022 ASEE Annual Conference and Exposition*. Minneapolis, MN.

Kirk, G. R., & Grohs, J. R. (2016). Civic attitudes and the undergraduate experience. In K. M. Soria & T. D. Mitchell (Eds.), *Civic engagement and community service at research universities: Engaging undergraduates for social justice, social change and responsible citizenship* (pp. 125–141). Palgrave Macmillan UK.

Knight, D. B. (2014). Reversing the logic: An outcomes-based student typology for determining "what works" in promoting an array of engineering-related student learning outcomes. *Educational Evaluation and Policy Analysis, 36*(2), 145–169.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037//0022-3514.77.6.1121

Lattuca, L. R., Knight, D. B., & Bergom, I. (2013). Developing a measure of interdisciplinary competence. *International Journal of Engineering Education, 29*(3), 726–739.

Lucia, A. D., & Lepsinger, R. (1999). *The art and science of competency models: Pinpointing critical success factors in organizations*. Jossey-Bass.

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, Albuquerue, NM.

Mazzurco, A., & Daniel, S. (2020). Socio-technical thinking of students and practitioners in the context of humanitarian engineering. *Journal of Engineering Education, 109*(2), 243–261. https://doi.org/10.1002/jee.20307

Meadows, D. H. (2008). *Thinking in systems: A primer*. Chelsea Green Publishing.

Miller, A. L. (2012). Investigating social desirability bias in student self-report surveys. *Educational Research Quarterly, 36*(1), 30–47.

Moely, B. E., Mercer, S. H., Ilustre, V., Miron, D., & McFarland, M. (2002). Psychometric properties and correlates of the Civic Attitudes and Skills Questionnaire (CASQ): A measure of students' attitudes related to service-learning. *Michigan Journal of Community Service Learning, 8*(2), 15–26.

Moore, S. M., Dolansky, M. A., Singh, M., Palmieri, P., & Alemi, F. (2010). *The systems thinking scale*.

Mosyjowski, E., Espinoza von Bischhoffshausen, J., Lattuca, L., & Daly, S. (2020). Student and practitioner approaches to systems thinking: Integrating technical and contextual considerations. In *2020 ASEE Virtual Annual Conference Content Access Proceedings*, 35219. https://doi.org/10.18260/1-2--35219

National Academy of Engineering. (2004). *The engineer of 2020: Visions of engineering in the new century*. National Academies Press. http://nap.edu/10999

National Academy of Sciences. (2004). *Facilitating interdisciplinary research*. National Academies Press.

National Institutes of Health. (2006). *Summary of the President's FY 2006 budget*. National Institutes of Health.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education, 15*, 624–632. https://doi.org/10.1007/s10459-010-9222-y

Novoselich, B. J., & Knight, D. B. (2017). Curricular and co-curricular influences on undergraduate engineering student leadership. *Journal of Engineering Education, 106*(1), 44–70. https://doi.org/10.1002/jee.20153

Paul, R. (1993). The logic of creative and critical thinking. In J. Wilsen & A. J. A. Binker (Eds.), *Critical thinking: How to prepare students for a rapidly changing world* (pp. 195–215). Foundation for Critical Thinking.

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

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education, 35*(1), 45–76. https://doi.org/10.1353/rhe.2011.0034

Porter, S. R. (2013). Self-reported learning gains: A theory and test of college student survey response. *Research in Higher Education, 54*, 201–226. https://doi.org/10.1007/s11162-012-9277-0

Randers, J. (2019). The great challenge for system dynamics on the path forward: Implementation and real impact. *System Dynamics Review, 35*(1), 19–24. https://doi.org/10.1002/sdr.1623

Reeb, R. N., Folger, S. F., Langsner, S., Ryan, C., & Crouse, J. (2010). Self-efficacy in service-learning community action research: Theory,

research, and practice. *American Journal of Community Psychology, 46*(3–4), 459–471. https://doi.org/10.1007/s10464-010-9342-9

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

Ro, H. K., Merson, D., Lattuca, L. R., & Terenzini, P. T. (2015). Validity of the contextual competence scale for engineering students. *Journal of Engineering Education, 104*(1), 35–54. https://doi.org/10.1002/jee.20062

Rosen, J. A., Porter, S. R., & Rogers, J. (2017). Understanding student self-reports of academic performance and course-taking behavior. *AERA Open, 3*(2), 1–14. https://doi.org/10.1177/2332858417711427

Senge, P. M. (2006). *The fifth discipline: The art and practice of the learning organization*. Crown Publishing.

Sosu, E. M. (2013). The development and psychometric validation of a Critical Thinking Disposition Scale. *Thinking Skills and Creativity, 9*, 107–119. https://doi.org/10.1016/j.tsc.2012.09.002

Steenkamp, J. B. E., De Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research, 47*(2), 199–214.

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

Sterman, J. D. (2018). System dynamics at sixty: The path forward. *System Dynamics Review, 34*(1–2), 5–47. https://doi.org/10.1002/sdr.1601

Svanström, M., Lozano-Garcia, F. J., & Rowe, D. (2008). Learning outcomes for sustainable development in higher education. *International Journal of Sustainability in Higher Education, 9*(3), 339–351.

Warburton, K. (2003). Deep learning and education for sustainability. *International Journal of Sustainability in Higher Education, 4*(1), 44–56.

Wheatley, M. J. (2005). *Finding our way: Leadership for an uncertain time*. Berrett-Koehler Publishers.

Wiek, A., Withycombe, L., Redman, C., & Mills, S. B. (2011). Moving forward on competence in sustainability research and problem solving. *Environment Magazine, 53*(2), 3–12.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.