Developing and Validating a Measurement of Systems Thinking: The Systems Thinking and Metacognitive Inventory (STMI). In, Routledge Handbook of Systems Thinking

Preprint · January 2021

DOI: 10.13140/RG.2.2.28639.38566

3 authors:

Laura Cabrera
Cornell University

Jessica Sokolow
Cornell University

Derek Cabrera
Cornell University

This is a draft preprint chapter from a forthcoming manuscript. Citation information as follows: Cabrera, L., Sokolow, J., and Cabrera, D. (2021) Developing and Validating a Measurement of Systems Thinking: The Systems Thinking and Metacognitive Inventory (STMI). In, Routledge Handbook of Systems Thinking, (Eds) Cabrera, D., Cabrera, L. and Midgley, G. Routledge. London, UK.

Developing and Validating a Measurement of Systems Thinking: The Systems Thinking and Metacognitive Inventory (STMI)

Laura Cabrera[12], Jessica Sokolow[12], and Derek Cabrera[12]

[1] Cornell University [2] CabreraResearch Lab

Abstract: There are a wide variety of psychographic and edumetric tests but none capture the totality of what it means to be a systems thinker. At this time systems thinking not only has popularity but also promises to address the complex problems we face today. It is therefore necessary to identify how to measure systems thinking on a continuum of mastery. The advent of DSRP Theory makes the variables that need to be measured explicit. We present herein a brief review of psychographic and edumetric advances in cognition, metacognition, and systems thinking, and then detail the development and validation of the Systems Thinking and Metacognition Inventory (STMI).

Measuring Systems Thinking Skills

The world we live in is increasingly complex, fast-moving and inundated with information. This new environment makes solving problems more complex. However, the greatest challenges are not derived from the complexity. In fact, they often result from a disconnect between how systems actually work in the real world, and our mental models of how they work. Systems thinking can help to reveal the underlying structure of systems and thought, regardless of the topic at hand [1]. The field of systems thinking was revolutionized with the discovery of four underlying cognitive skills[1] that are present in all systems thinking frameworks. These underlying skills are: making distinctions, organizing part-whole systems, recognizing relationships, and taking perspectives (DSRP) [1–6]. These rules can help reveal the underlying structural patterns of our thinking, revealing the pitfalls that complicate many of our world's systems.

The simple rules of DSRP are intrinsic to how we structure and organize information to make meaning (i.e., how we think). Applying these rules consciously and explicitly brings awareness to how we structure and organize that information, i.e., thinking about thinking or metacognition. This process of metacognition is important to ensure that, for example, we not only address a perspective, but we do so consciously and ensure we do not omit other key perspectives. DSRP's application provides a process for questioning accepted mental models that impact the way we see the world, and which we build using accepted ideas and perspectives. The simple, universal and thus democratic nature of these simple rules makes DSRP applicable to everyone.

There is a body of research (e.g., [7,8]) that calls upon the benefits of systems thinking education, but questions still remain about the effectiveness of educational approaches. This is due in part to the fact that there lacks a method to measure systems thinking abilities in individuals. And this is not an easy task, as it requires understanding and measuring individuals cognitive and metacognitive abilities. However, the discovery of DSRP—universal cognitive structures that underlie systems thinking—means that new efforts to measure systems thinking may be more feasible than ever before.

The *systems thinking and metacognitive inventory* (STMI) aims to address this assessment gap. The STMI measures individual mental capabilities and processes to think systemically and think about one's thinking (i.e., metacognition). It also measures perceived skills and provides a comparison to baseline cognitive/metacognitive skills. The STMI serves as a tool and resource to help educators, institutions and individuals track the progress in developing systems thinking skills. Previous efforts have been made to measure the systems thinking capabilities of individuals, as will be discussed briefly in this paper [9–11]. However, many attempts are often discipline-specific (e.g., [10]), or rely solely on largely subjective qualitative analyses [9]. While the latter tests have the potential to reveal fascinating emergent results, they lack scientific validity, and the ability to objectively measure progress and to scale assessments. Furthermore, existing tests measure aspects of systems thinking theories, but fail to measure the cognitive abilities that underlie all forms of systems thinking (or DSRP). STMI serves as an easily

[1] In this paper, we will call the four DSRP patterns/rules 'skills,' because we are measuring DSRP in humans. DSRP are patterns found in mind and nature. They exist, regardless of whether anyone develops or practices them. When they are measured for the purpose of explicating them and transforming them into metacognitive tools, they become 'skills.'

reproducible measurement tool that consumes less time, resources and is less biased compared to other methods of testing skill uptake, such as self-assessments, qualitative interviews or observations.

Edumetric Testing and Its Origins

The STMI is an *edumetric* test, measuring the ability of an individual as a snapshot in time, which then can be replicated over time. This is different from a *psychometric* test (e.g., the Myers Briggs Type Indicator), which are fixed measures that intend to be a diagnosis of an individual's cognitive or behavioral abilities or habits [12]. Psychometric tests focus on 'individual difference,' whereas edumetric tests focus on 'individual gain or growth' [12] Because one's ability with systems thinking and metacognition is something that can be learned or improved, the focus of STMI is on growth or gain and is therefore edumetric.

Edumetric testing, while distinct, originates from psychometrics "a scientific discipline concerned with the question of how psychological constructs (e.g., intelligence, neuroticism, or depression) can be optimally related to observables (e.g., outcomes of psychological tests, genetic profiles, neuroscientific information)" [13]. The foundations of psychometric testing date back to Charles Darwin's *The Origin of Species* [14] in 1859, which explained why individual members of the animal kingdom differ. In his works, he documented the importance of adaptive traits to a species' survival, where these traits are passed to successive generations. Later, Sir Francis Galton, considered the 'father of psychometrics' sought to measure these adaptive traits introduced by his half-cousin, Darwin. Galton established the Anthropometric Laboratory and devoted much of his career to the statistical analysis of humans physical and mental traits [13,15]. Galton's early approaches to measuring individual differences, included sharpness of sight through performance accuracy and reaction times [13].

Later, interest in the field grew as the United States Military implemented a program to select prospective soldiers by measuring psychometric abilities relevant to performance. From this program arose two important aspects of psychometric testing: reliability and validity. First, it was important the measurement instrument would yield the same outcomes when measuring an individual with the same level of standing of the attribute being measured, under the same circumstances (reliability). Second, it was important that the measurement instrument was valid—that is, it measured *only* what it intended to measure [13]. These two components remain important aspects of determining the quality of both psychometric and edumetric testing today.

Psychometric Measures of Cognition versus Personality

Due to the fact that many of the measures related to systems thinking and DSRP developed thus far are psychometric measures, we will provide some background on psychometric measures as context for edumetric measures. To understand edumetric testing, it is helpful to also understand psychometric testing.

Psychometric testing contains two primary flavors: *mental ability and aptitude ('cognitive') testing* and *personality testing*. We will review both in this section, but focus more on cognitive, as these types of tests are more aligned with the content and purpose of STMI.

One of the most widely known personality tests is the Meyers Briggs test (or the MBTI). The MBTI was developed from a Jungian theory of personality types but also over decades of work with diverse samples [16]. The single most remarkable thing about the test seems to be its popularity. MBTI's validity, however, is questionable [17,18]:

> In 1991, the National Academy of Sciences committee reviewed data from MBTI research studies and concluded that only the I-E scale has adequate construct validity in terms of showing high correlations with comparable scales of other instruments and low correlations with instruments designed to assess different concepts. In contrast, the S-N and T-F scales show relatively weak validity. The 1991 review committee concluded at the time there was "not sufficient, well-designed research to justify the use of the MBTI in career counseling programs".

Overall, the NAS review committee concluded that the MBTI has not demonstrated adequate validity (nor reliability) to be used in career counseling, despite its growing popularity for that purpose.

Psychometric tests that assess ability and aptitude (including intelligence and cognitive ability) are most closely related to the STMI. Raymond Cattell suggested that cognitive tests assess fluid and/or crystallized intelligence, where fluid intelligence refers to one's general abilities to think or learn and crystallized intelligence refers to the results of fluid intelligence (e.g., thinking abilities vs. domain-specific knowledge; [19–21]). For this paper, we are particularly concerned with metacognitive and cognitive abilities, due to the role systems thinking has in fostering cognitive and metacognitive functions.

Measuring Cognition

Perhaps the best known, and most commonly administered [22], modern psychometric measure assessing cognitive ability is the Binet–Simon IQ (or intelligence quotient) test. The test was utilized for the first time by the French government in the late nineteenth century. Binet and Simon sought to develop a baseline of intelligence by testing a wide range of children across a multitude of measures. However, they failed in developing one clear baseline indicator of intelligence and therefore evolved their approach to ascertain common levels of intelligence across categories and age groups [23]. The IQ test is designed to measure an individual's general ability to solve problems and understand concepts in relation to the rest of the population. The IQ test is constructed based on the ratio of mental age to chronological age, multiplied by 100. The reliability of IQ tests generally falls between .80 and .90 [24].

The validity of IQ tests derives largely from their predictive power. Kaufman writes, "As a measure of past learning, the IQ test is best thought of as a kind of achievement test, not as a simple measure of aptitude" ([25], pg. 25). As Braden and Niebling ([26], p. 740) write, "Intelligence tests are more often criticized for failing to assess their intended construct (intelligence) than for lacking reliability…" Meta-analyses have shown that IQ is highly predictive of training success and job performance. There have been thousands of studies, many with high numbers of respondents. IQ is also strongly correlated with occupation, years of schooling, academic achievement, etc. There are 10 commonly used IQ tests, with the Wechsler

Adult Intelligence Scale (WAIS) seeming to be favored for adults. Test scores from different IQ tests tend to correlate highly. It is disputed that IQ tests measure general cognitive ability (e.g., [27]).

An alternative psychometric test of intelligence to the IQ test is the Differential Ability Scales (DAS). The DAS consists of six subtests, testing verbal and non-verbal abilities and spatial abilities: Word Definition, Similarities, Matrices, Sequential and Quantitative Reasoning, Recall of Design, and Pattern Construction [28]. The Cronbach alpha coefficients of internal consistency ranges from .79 to .91 for all age groups, and the reliability ranges from .65 to .90 for two age groups [29]. Studies comparing the subtests to other cognitive assessments result in relatively high correlation coefficients, thus demonstrating relatively high construct validity (approximately .51–.87; [28]).

Other tests of cognitive ability have been developed, beyond the IQ test. These tests are typically psychometric assessments and are widely used by institutions to measure general intelligence (such as IQ) and competence and suitability, particularly related to job performance. In fact, general cognitive ability (often used synonymously in the literature with intelligence) is a valid predictor of job performance [30].

Measuring Creative Thinking

A central goal in education and the workplace is to get individuals to think more critically and creatively. We not only want people to think more deeply about complex problems, but we want to offer novel thoughts and solutions. There is some disagreement among scholars and theorists about the relationship between critical and creative thought. Some believe they are different but complementary, while others believe they are in conflict [31]. Regardless, researchers generally agree that the two are pedagogically different [31], and thus require different developmental techniques as well as measures.

Creativity is an important aspect of our mental process, and involves cognitive and metacognitive processes to develop creative products or solutions. Assessments of creativity can help to understand these processes, supporting the development of more creative thinkers [32]. Hocevar [33–35] identified four tests of creativity in: divergent thinking tests, attitude and interest inventories, personality inventories, and biographical [32]. Other measures include self-ratings [36–40] and expert-ratings [39] of creativity. Some scholars defend the social nature of creativity, and the role of social validation in instances of creativity [41]. Based on this understanding, expert-ratings are a useful assessment to determine the societal recognition of a creative product [39].

Divergent thinking tests, in which a range of answers (e.g., open-ended solutions) may be considered correct, are the most common methods to assess creative or out-of-the-box thinking [39]. Divergent thinking often leads to originality, which is central to creativity [42]. This method contrasts with convergent thinking, which aims to use available information to determine a single solution to a problem [43]. The predictive and discriminant validity of divergent thinking tests enjoys mixed support [32,41]. This validity, of course, varies under the myriad of sampling and assessment conditions, which impacts the methodological rigor of the research [44].

Although our focus herein is not to document all the tests that measure creative thinking, in order to provide some ‘flavor’ for the types of tests that exist, we have included a few of the most widely cited psychometric instruments.

Structure of the Intellect Inventory

The Structure of Intellect (SOI) Inventory is based on Gildford’s [45] model of creativity tests individual creativing related to semantic systems (e.g., thinking of consequences if people no longer need to sleep), figural systems (e.g., constructing meaningful figures from sets of elements), and symbolic units (relating a set of numbers in different ways to get a certain result. The test scoring is based on the four criteria: fluency (the number of relevant ideas), flexibility (the number of categories which the responses fall into), originality (the number of unusual answers determined by relative statistical frequency), and elaboration (the number of details for a response; [32]).

Torrance Test of Creative Thinking

This test is the most widely used and accepted measure of creativity [32,44,46–48].The Torrance Test of Creative Thinking (or TTCT) was developed by E.P. Torrance [49], based on the Structure of Intellect (SOI) Divergent Production test, and captures individual creative potential and demographic variations [32]. Torrance [49] defines creativity as “a process of becoming sensitive to problems, deficiencies, gaps in knowledge, missing elements, disharmonies, and so on; identifying the difficulty; searching for solutions, making guesses, or formulating hypotheses about the deficiencies: testing and retesting these hypotheses and possibly modifying and retesting them; and finally communicating the results.” ([49], pg 6).

The form of the test is both verbal and figural (though these are not the only forms creative thinking can manifest; [32]). Verbal questions include five activities: asking questions to understand what is happening in a drawing or guessing possible causes that lead to action shown in a drawing, product improvement, unusual uses, unusual questions, and just suppose. Figural questions consist of three activities, completing and designing a picture, and repeated figures of lines or circles. The scoring criteria differs between the figural and verbal tests. The figural test is based on originality, elaboration, task performance, fluency and flexibility [47]. The scoring of the verbal test is based on fluency, flexibility, and originality. Based on the 1966 and 1974 TTCT Manual, the test-retest reliability coefficients range from .50 to .93, which was considered reasonable by Torrance [50] and Treffinger [51]. The reliability score is considered While the predictive validity of TTCT is questioned as a divergent test, empirical evidence 22-year follow-up study of elementary school students indicated a predictive validity of 0.63. A 40-year follow-up study from the same sample of students found that TTCT explained 23% of the variance in creativity [32].

Wallach–Kogan and Getzels–Jackson Tests of Creative Thinking

The Wallach–Kogan Creativity Tests (WKCT) and Getzels–Jackson Creativity Tests (GJCT) are similar to the SOI inventory and the TTCT. WKCT is based on the associative conception of creativity, where associated elements form combinations to meet stated requirements or have some utility [52]. WKCT activities include verbal activities (e.g., thinking possible uses for an object and figural activities (e.g., determining a line drawing's meaning). The test is scored based on the number and uniqueness of associational responses [53].

The Getzels–Jackson test (1962) includes five sections: 1) word association (generating definitions to words), 2) uses for things, 3) hidden shapes (finding the geometric figure hidden in a more complex pattern), 4) fables (composing different endings for a fable), and 5) make-up problems (proposing math problems in a context). Scores of these tests are also based on fluency, flexibility, originality, and elaboration [32].

Other psychometric measures of creativity have been developed, including the Creative Scientific Ability test (C-SAT). The test measures one's potential for scientific creativity in regards to hypothesis generation, experiment design and evidence evaluation. The authors found that the internal consistency was found to be good at .87 and the inter-scorer reliability at .92 [54].

Measuring Critical Thinking

Critical thinking has been studied in detail to better understand human cognition, although there is no universally accepted definition. Attempting to capture the complex, multifaceted nature of the concept [55], the American Philosophical Association, defines critical thinking as:

> "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based" ([56], p. 2).

Bonnefon [57] argues that this definition aligns with dual processing theory of thinking and reasoning. Bonnefon [57] and Halpern [58] equate critical thinking with Type 2, or analytical processes that are purposeful, self-regulatory, conscious and effortful processing as opposed to Type 1, or fast, implicit, automatic processes (i.e., intuition) [55]. Type 2 processing is associated with working memory capacity and executive function [59], but also with dispositions, attitudes, beliefs and motivation (e.g., [60]). Dual process theories reflect the importance of both reflective thinking and metacognitive processes [55].

Critical thinking skills refer to "rules of formal logic, consideration of multiple perspectives, including induction and deduction" [61]. There are different methods of testing these skills, including methods of forced-choice and constructed-response questions, where the latter allows assessors to infer the respondent's reasoning for an answer. Examples include [62]: "the HCTA is a test that combines constructed-response and forced-choice items. Apart from the above mentioned item formats, still other formats have been used, such as interviews (e.g., the Reflective Judgement Interview), essays (e.g., Ennis-Weir Critical Thinking Essay Test), a combination of essays and multiple-choice questions (e.g., the Critical Thinking Assessment Battery), and Likert-type statements (e.g., the Problem Solving Inventory)."

When assessing critical thinking, it is important to have a clear sense of the purpose of the test. There are certain tests for certain levels of assessment. The most common critical assessments that nearly every college-educated individual has engaged include: ACT, SAT, AP, GRE, MCAT, or LSAT tests. While these are examples of high-stakes critical thinking tests. There are other validated tests that cover more than one area of critical thinking. As Verburgh explains, "the diversity of conceptualisations of critical thinking are mirrored in a diversity of available discipline-specific and discipline-general tests of critical thinking" [62].

Common Critical Thinking Tests

Models for different assessments of CT reflect the different understandings of what elements encompass critical thinking, as well as the applicable disciplines. A number of generic measures have been designed to assess critical thinking skills. These measures tend to focus on problem solving in critical thinking [55].

There are a variety of methods used to measure critical thinking, including forced-choice and constructed-response items, where the latter "allows researchers to infer the respondent's reasoning behind an answer. Most researchers agree that there are some general CT skills, which are applicable in various contexts, while familiarity with a discipline plays an important role too" [61]. Ennis is a leading figure in the generalist movement, where CT is believe to be "a set of cognitive abilities that can be independently taught of a specific content" [62]. Comprehensive analyses of available tests are available in literature reviews, such as Ennis [63].

Liu et al., [61], explains that most validated tests use fixed response question types: "The majority of the assessments exclusively use selected-response items such as multiple-choice or Likert-type items (e.g., CAAP, CCTST, and WGCTA). Given the limited testing time, only a small number of constructed-response items can typically be used in a given assessment." And that, "Although constructed-response items have great face validity and have the potential to offer authentic contexts in assessments, they tend to have lower levels of reliability than multiple-choice items for the same amount of testing time." Liu et al., [61], also explains that, "Critical thinking assessments showed moderate correlations with general cognitive assessments such as SAT® or GRE® tests."

Liu et al., [61], provides two tables on the major critical thinking tests. The first table [p.5-7], entitled, "Existing Assessments of Critical Thinking," explores the various critical thinking tests, vendors, formats, delivery types, length, 'forms and items' (number of questions and their types), and 'themes/topics' (e.g., scales, etc.). The second table [p.9-10], entitled, "Validity Evidence," explores various studies of critical thinking assessments, the subjects in those studies, sample size, and validity of the tests. We do not reproduce those tables herein, but suggest the reader consults these tables. Below, we reproduce Liu et. al.'s critical thinking tests with basic information about their 'themes/topics' as per Lui et. al. [61] to provide the reader with the names and specific of what these tests measure.

California Critical Thinking Disposition Inventory (CCTDI)

"This test contains seven scales of critical thinking: (a) truth-seeking, (b) open-mindedness, (c) analyticity, (d) systematicity, (e) confidence in reasoning, (f) inquisitiveness, and (g) maturity of judgment" [61].

California Critical Thinking Skills Test (CCTST)
"The CCTST returns scores on the following scales: (a) analysis, (b) evaluation, (c) inference, (d) deduction, (e) induction, and (f) overall reasoning skills" [61].

California Measure of Mental Motivation (CM3)
"This assessment measures and reports scores on the following areas: (a) learning orientation, (b) creative problem solving, (c) cognitive integrity, (d) scholarly rigor, and (e) technological orientation" [61].

Collegiate Assessment of Academic Proficiency (CAAP)
"The CAAP Critical Thinking measures students' skills in analyzing elements of an argument, evaluating an argument, and extending arguments" [61].

Collegiate Learning Assessment + (CLA +)
"The CLA +PTs measure higher order skills including: (a) analysis and problem solving, (b) writing effectiveness, and (c) writing mechanics. The MC items assess (a) scientific and quantitative reasoning, (b) critical reading and evaluation, and (c) critiquing an argument (Zahner, 2013)" [61].

Cornell Critical Thinking Test (CCTT)
"Level X is intended for students in Grades 5–12 + and measures the following skills: (a) induction, (b) deduction, (c) credibility, and (d) identification of assumptions (The Critical Thinking Co., 2014) Level Z: 52 items Level Z is intended for students in Grades 11–12 + and measures the following skills: (a) induction, (b) deduction, (c) credibility, (d) identification of assumptions, (e) semantics, (f) definition, and (g) prediction in planning experiments" [61].

Ennis–Weir Critical Thinking Essay Test
"This assessment measures the following areas of the critical thinking competence: (a) getting the point, (b) seeing reasons and assumptions, (c) stating one's point, (d) offering good reasons, (e) seeing other possibilities, and (f) responding appropriately to and/or avoiding argument weaknesses" [61].

ETS Proficiency Profile (EPP)
"The Critical Thinking component of this test measures a students' ability to: (a) distinguish between rhetoric and argumentation in a piece of nonfiction prose, (b) recognize assumptions and the best hypothesis to account for information presented, (c) infer and interpret a relationship between variables, and (d) draw valid conclusions based on information presented" [61].

Halpern Critical Thinking Assessment (HCTA)
"This test measures five critical thinking subskills: (a) verbal reasoning skills, (b) argument and analysis skills, (c) skills in thinking as hypothesis testing, (d) using likelihood and uncertainty, and (e) decision-making and problem-solving skills" [61].

Watson–Glaser Critical Thinking Appraisal tool (WGCTA)
"The WGCTA is composed of five tests: (a) inference, (b) recognition of assumptions, (c) deduction, (d) interpretation, and (e) evaluation of arguments. Each test contains both neutral and

controversial reading passages and scenarios encountered at work, in the classroom, and in the media. Although there are five tests, only the total score is reported" [61].

Measuring Metacognition

An important aspect of systems thinking is the act of metacognition. The process of deliberately structuring one's thoughts using the four building blocks of cognition (D, S, R and P) requires awareness of, or thinking about, one's own thinking, or metacognition.

The term *metacognition* was first used by Jon Flavell in 1979 to understand the distinction "between knowledge about the contents of memory versus processes used to regulate and monitor memory and cognition" ([64] p. 34). Metacognition is often defined as "the activity of monitoring and controlling one's cognition" [65]. Research on metacognition distinguishes between two major components: knowledge of cognition (i.e., an individual's awareness of their own metacognitive process) and regulation of cognition (i.e., an individual's ability to control and/or direct their metacognitive process) [64,66,67]. One's knowledge of cognition includes three distinct sub-processes [67]:

- Declarative Knowledge, i.e., an individual's understanding of their intellectual skills, resources, and abilities
- Procedural Knowledge, i.e., an individual's understanding of how to implement learning strategies; and
- Conditional Knowledge, i.e., an individual's understanding about when and why to use learning procedures.

Regulation of cognition measures an individual's understanding of their knowledge about "planning, comprehension monitoring, and evaluation" [67]. Not all of these sub processes result in explicit, clearly defined behavior [68]. Further, very often individuals are not aware of the process that drives metacognition.

Researchers find that there currently exists no single method that measures all metacognitive processes [68,69]. Measurement tools that are used to measure metacognition are often based: 1) on an individual's own telling (questionnaires and interviews) or 2) objective behavior measurements (i.e. systematic observation and think aloud protocols). Both approaches take considerable time, as they require each respondent to be evaluated individually. Akturk and Sahin [68] reported on the benefits and drawbacks to these approaches. Think aloud protocols require subjects to state verbally how they will handle a problem. In a learning environment, this process can result in drawbacks, including: requiring students to leave their learning environment and also preventing students from learning the material. Interviews are useful in that they enable an in-depth investigation of students' ideas, however they require a great deal of time and yield a limited number of responses. The think aloud and interview approaches allow the researcher to understand otherwise unobservable processes, and can be useful for initial investigation. However, challenges of inadvertent cuing, memory failure and other interview biases can prevent the accuracy of this method [70]. Questionnaires allow researchers to evaluate a large group of students at one time, and thus is a faster, more objective approach than interviews or think aloud protocol. However, subjects may be reluctant to express their ideas and experiences and may choose more socially attractive responses, thus leading to bias [68,70].

The most common questionnaire, in the form of a psychometric assessment, is the Metacognitive Assessment Inventory (MAI) [67]. The MAI focuses on two aspects of metacognition and the associated sub-processes, regulation of cognition and knowledge of cognition, as well as an overall assessment of cognition, based on the theoretical foundations [65]. The MAI is a 52-item test, and is suitable for adolescents and adults [67]. The 1994 study by Schraw and Dennison found that the internal consistency of these scales was high, ranging from .93 to .88. There was a high degree of similarity between factor loadings across the experiments performed [67]. The MAI correlated with reading comprehension test performance only on the knowledge of cognition factor of the test [65]. Other assessments are not as positive, for example Sperling et al [71] performed an assessment of the MAI in relation to measures of academic achievement, other than reading comprehension, including the SAT and high school GPA, and did not find correlation [65].

Measuring Emotional Intelligence

Improvement in systems thinking ability is found to improve an individual's overall emotional intelligence [72]. Emotional intelligence abilities are typically understood by individuals "to apply information provided by emotions for the improvement of cognitive processing" [73]. Emotional intelligence is the combination of mental abilities, stable behavioral traits and personality variables." [73].

There is a distinction between *trait* and *ability* in emotional intelligence. Perez and colleagues [74] explain:

> "...Petrides and Furnham (2000a, 2000b, 2001) distinguished between trait EI (or emotional self-efficacy) and ability EI (or cognitive-emotional ability). It is important to understand that trait EI and ability EI are two different constructs. The former is measured through self-report questionnaires, whereas the latter ought to be measured through tests of maximal performance. This measurement distinction has far-reaching theoretical and practical implications. For example, trait EI would not be expected to correlate strongly with measures of general cognitive ability (*g*) or proxies thereof, whereas ability EI should be unequivocally related to such measures."

There are both trait (self-report) and ability tests of emotional intelligence (EI). The original tests of EI were self-report. Brackett, Delaney, and Salovey [75] note that self-report assessments, or surveys asking respondents to report on their own emotional skills are:

> "most often associated with mixed and trait models. Self-report measures are usually quick to administer. However, many researchers argue that their vulnerability to social-desirability biases and faking are problematic. In addition, there is wide speculation concerning the potential for inaccurate judgments of personal ability and skill on behalf of responders. Self-report measures have been shown to lack discriminant validity from existing personality measures and have very low correlations with ability measures of EI."

The primary ability tests of emotional intelligence were developed by Mayer and his colleagues. Tests are available to assess emotional perception [76], but this is at a more narrow scope that does not assess overall emotional intelligence [77]. Mayer and colleagues' developed two influential tests, the Multifactor Emotional Intelligence Scale (MEIS) [78], and its successor, the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) [79]. Both tests measure the conceptualization of EI as 'four branches:' "emotional perception, emotional facilitation of thought, emotional understanding, and emotional management," with 141 items divided into 8 tasks with the MSCEIT in the four branches with 2 areas scores, and a total score (i.e., 7 scores)[80]. Brackett and Mayer ([81]) describe the MSCEIT as both reliable and content valid insomuch as, "reliability for the second version of MSCEIT is reliable at the full-scale level (r's = .90 to .96), the area level (r's = .84 to .91), and the branch level (r's = .74 to .91)" [80]. The test is also found to be predictive for verbal SAT scores [74].

Much of this review is simply to establish what has come before in terms of cognitive testing, the types of measures that have been constructed, and the results. This provides a backdrop for the motivation and steps involved in measuring system thinking.

Measuring Systems Thinking-DSRP

DSRP Theory provides a universal foundation upon which systems thinking frameworks of all types are based [1,2,5,72,82–87]. The evidence-base for DSRP Theory is extensive [2,83,85,88] (see Chapter 35 in this volume) in the systems sciences, systems thinking, cognitive science, and neuroscience literatures, as well as across the disciplines from physics to sociology. A summary of a vast literature is often difficult, if not inadvisable. This holds true for systems thinking, which is of interest to scientists and academics, as well as practitioners and the general public. For this reason, we have tried to summarize the literature in a way that is accessible *and* maintains fidelity to what the literature says. A summary of the transdisciplinary literature relating to systems thinking points to the following conclusions:

1. There is a *universality* of certain structures (DSRP), namely:
    a. Identity-Other Distinctions — D(*i*⇄*o*)
    b. Part-Whole Systems — S(*p*⇄*w*)
    c. Action-Reaction Relationships — R(*a*⇄*r*)
    d. Point-View Perspectives — P(*ρ*⇄*v*)
2. These DSRP structures exist in both mind (cognition) and nature (a.k.a., reality);
3. Awareness of these structures (a.k.a., metacognition) makes a significant difference in all walks of life, and success in all domains; and
4. Systems thinking is an emergent property of the DSRP simple rules.

Cabrera [6] defines a mental model (𝕄) as the complex product (⊗) of information (𝕀) and DSRP simple structural rules of thinking (𝕋) shown in Table 1. Thus the definition of a mental model: 𝕄 = 𝕀 ⊗𝕋.

The Identity-Other Distinctions Rule	
D := (*i*↔*o*)	A Distinction (D) is defined as *identity* (i) co-implying an *other* (*o*)

The Part-Whole Systems Rule	
$S := (p \leftrightarrow w)$	A System (S) is defined as *part* (*p)* co-implying a *whole* (*w*)
The Action-Reaction Relationships Rule	
$R := (a \leftrightarrow r)$	A Relationship (R) is defined as *action* (*a*) co-implying a *reaction*(*r*)
The Point-View Perspectives Rule	
P := (*ꝑ*↔*v*)	A Perspective (P) is defined as *point* (*ꝑ*) co-implying a *view* (*v*)

Table 1: The universal structural simple rules of thought [1].

It is increasingly clear that these D, S, R, and P structures cannot be separated from each other. In other words, they are in a constant and dynamic interplay. It is clear that each of the D, S, R, and P structures is *necessary* and *sufficient* for each of the individual structures to exist. This means that, in order for an S (part-whole system) to be formed, Ds (distinctions) need to be made, Rs (relationships) need to exist, and Ps (perspectives) need to be taken; and it is the same for the others—each one requires the other three. In addition, any given D, S, R, and P structure cannot be divorced from its two elements, and nor can the elements be divorced from each other. For example, we are near-constantly making identity distinctions (D*i*), and more often than not unaware of the *other* distinctions (D*o*) we are implying. Nevertheless, they exist. Said another way, it is impossible *not* to make distinctions, connections, groupings, and framings.

That DSRP structures exist in both mind (cognition) and nature (a.k.a., reality) is perhaps read as a philosophical statement, but it is based on *empirical* findings [2,83,85,88]. It does not mean that every DSRP structure that we build in our minds exists in nature or vice versa. It means that the D, S, R, and P structures are found in both realms. This is important because it tells us something about one of the goals of systems thinking: to increase the *probability* that our mental models are in *alignment* with reality. Working with a set of structures that is common to both mind and nature—akin to finding the same denominator to add fractions—is an essential first step in increasing the probability of alignment between our thoughts about reality and the actual reality in which they exist.

Two further well-known goals of systems thinking are also enhanced by DSRP: (i) exploring possibilities for addressing uncertainties, and/or (ii) taking future action: we can think about what will, can, could, would or might be, either now or if we undertake an intervention. The capability that enables this is *critical* thinking, or what Cabrera calls ‘structural prediction’ [1,89–92] (also see Chapter 7): we can rethink boundary Distinctions, explore different part-whole System configurations, take different Perspectives, and look for possible new Relationships. This kind of critical thinking generally has to be anchored in an alignment with reality, as the conditions surrounding a given uncertainty might be relevant to addressing it, and action for change has to take us from the current situation to a new one, so knowing about that current situation matters.

Finally, *deliberate departures* from alignment with reality can produce creative fantasies, ranging from scenarios that are just a small plot twist away from our current status quo, to totally surreal worlds that are hilarious in their absurdity. The creation of 'what if' scenarios or creative thinking can be aided by DSRP. What if we Distinguished a universe existing within each living cell? What if our known universe was just a part of a larger System with different physical laws? What if it were possible for an altered mind to establish Relationships with the minds of others and exert control over them? What if we took an unusual Perspective, such as that of an insect living in our house? It is clear that DSRP can underpin the creation of fiction as well as fact. Given strong claims in the fields of systems thinking and systems science that what we do is transdisciplinary (i.e., using ideas that work across disciplinary boundaries), it makes sense to us that this should not only work for the sciences (e.g., [93]), but can extend into the arts too.

Awareness of these DSRP structures constitutes 'systems thinking skills'. *Cognition* often occurs, whether we like it or not, in ways that we are unaware of. *Metacognition,* or the awareness of how we think (in D, S, R, P structures and dynamics), shows us the structural underpinnings of the mental models we build about the world. This structural understanding further allows for the identification of gaps in our knowledge, and provides insights into how to fill them. DSRP can be thought of as a set of metacognitive *skills* that can be taught, learned, and practiced [94–102]. The Systems Thinking and Metacognition Inventory [103,104] or STMI, is a validated and reliable edumetric test of these four skills. The STMI measures individual mental capabilities and processes to think systemically and think about one's thinking (i.e., metacognition). The STMI serves as a tool and resource to help educators, institutions, and individuals track the progress in developing systems thinking skills.

The idea (mentioned earlier in this section) that systems thinking is an emergent property (i.e., it is a Complex Adaptive System or CAS) [1,83,84,89,105,106] is a significant shift in pragmatic terms. It means that if one wants to 'do systems thinking' or 'get better at (developing the skill of) systems thinking,' one must recognize that systems thinking itself is an emergent property of a process. It is not something you can 'do' per se, but is rather something you 'get' when you do DSRP. Although 'doing DSRP' does not require metacognition (it is hypothetically possible for a person to be good at doing DSRP without knowing it), being aware of the DSRP structures (metacognition) significantly increases the probability of one 'doing DSRP' and therefore 'getting systems thinking.' It also means that [technically speaking] neither DSRP, nor the various frameworks launched in the first three waves, are 'what systems thinking *is*.' Systems thinking is a complex and adaptive system, an outcome—an emergent property—of the DSRP simple rules and information agents. Those multiple frameworks/methodologies from the first three waves might not *be* systems thinking, but they can usefully support the operationalization of the DSRP elements in complex combinations that are useful for different purposes (and the many visualization and modelling techniques that come with the frameworks likewise aid operationalization).

Existing Attempts to Measure of 'Systems Thinking'

According to Doyle [107] researchers have made many claims about the relationship between systems thinking and cognitive processes. Stave and Hopper [108] demonstrated through their meta-analysis of systems thinking intervention research that there is a breadth of qualitative evidence that systems thinking improves critical thinking. However, there is very limited quantitative research, through rigorous psychometric methods, to understand the impact of systems thinking on human cognition ([109]. The STMI serves as a tool to more deeply understand this relationship from the perspective of the individual learner. There is a breadth of understanding that there is a gap in regards to the availability of empirical evidence of the effectiveness of systems thinking (e.g., [110–112]). The following tools offer initial attempts to conceptualize how to evaluate and measure an individual's systems thinking abilities and skills.

Framework Systems Thinking Questions

Dorani et al [113] recognized that the structure of methods of assessing systems thinking abilities were impractical for testing large numbers of examinees [113]. Dorani and colleagues cite previous efforts to collect data through interviews or observational protocol (time intensive), self-assessments (biased), and skill-level examinations, which currently are not holistic assessment of system thinking abilities [113]. To address these concerns, Dorani and colleagues developed a set of guidelines to develop systems thinking questions for standardized assessments. Dorani and colleagues covered six topic areas, each covering a critical system thinking skills: dynamic thinking, system-as-cause thinking, forest (or holistic) thinking, operational thinking, closed-loop thinking, stock-and-flow thinking [114–116]. The proposed methodology is such that each question should be prefaced with a relevant scenario, in such that systems thinking skills are isolated from each other.

Links Between Systems Thinking and Complex Decision Making

Maani and Maharaj [110] developed the Verbal Protocol Analysis (VPA) methodology (or a think aloud protocol) to assess individual systems thinking skills. For the VPA methodology, participants verbalize their thought process to complete a task, in this case in regard to a business case to assess revenue, market share and profit [110] . The results of the VPA methodology were broken into fragments, which were then coded based on a 5-level classification of systems thinking categories.

Maani and Maharaj developed an assessment protocol to quantify and compare results from each respondent. To do this, the authors first developed categories to assess systems thinking abilities based on the theoretical foundations of Richmond's [114]. These five categories included (ranked from highest to lowest systems thinking abilities, based on the author's research and understanding) include: 1) forest thinking, 2) closed-loop thinking, 3) operational thinking, 4) system-as-cause thinking and 5) dynamic thinking. A cumulative notion was used, in which the higher ranking category is inclusive of each lower category (i.e., forest thinking includes all other categories). Based on these measures, a percent frequency measure of occurrence of each systems thinking type was developed to quantify the analysis and compare across respondents. For example, if a VPA contained 380 fragments, and 56 were coded with forest thinking, the frequency would be 56 divided by 380, or 14.74 percent.

In addition to the classification, Maani and Maharaj also developed a measure "to determine how well participants understood the structure of the task and the relationships" contained within the task system. Relationships were coded based on the following categories: basic one-to-one relationships (largely intuitive), complex one-to-one relationships, three-way relationships, and big picture. Each relationship obtained a score, as determined by the category in which it was contained (as pre-defined by the protocol), where basic relationships received the lowest points.

The study was tested among 10 business graduate students (ages 20 to 25), where all participants had systems thinking training. Maani and Maharaj explained in their paper that while their sample size was small, due to the large amount of information obtained from the VPA methodology, the sample size was adequate, explaining that typical VPA samples range between 10 and 20. Maani and Maharaj, or subsequent authors, did not report on the validity or reliability of the test, and its ability to draw conclusions about individual systems' thinking abilities.

Cognitive Mapping Assessment of Systems Thinking

Some researchers have worked to assess systems thinking abilities through the creation of concept maps. The Cognitive Mapping Assessment of Systems Thinking (CMAST) focuses "on the causal structures that students develop mentally in order to understand the dynamics of complex systems" [117]. CMAST measures deficiencies in causal maps used to understand the dynamics of the system in relation to the cause and effect relationships [117]. CMAST provides a measure of respondent's use of causal loops on a visual map used to understand and describe complex systems [117]. The assessment aims to measure how individuals organize new information, of which they have little to no previous knowledge. Plate [117] assesses these maps via link density, to measure differences in closed branching and causal loops, and a Web-link Causality Index (WCI), a calculation of the percentage of notes with more than one cause and/or effect. There have been other efforts to score concept maps [118,119]. These assessments are based on a scoring system that takes into consideration a number of metrics, such as: total number of connections among map concepts, number of concepts related to the theme represented by a map, etc. [118,119].

Systems Thinking Assessment Tool

Grohs' 2015 dissertation [9] offers a framework for evaluating an individual's systems thinking competency through a scenario-based assessment tool. Grohs conceptualizes "systems thinking" as a metacognitive strategy that offers a flexible way of "framing, reasoning, and acting within multiple dimensions ([9] p. 56). This assessment tool is based on a conceptual framework, developed by Grohs based on [his] review of systems thinking literature from multiple disciplines. The framework considers three main distinctions:

1. Time (reflective ability, predictive ability),
2. Social (stakeholder identification, incorporation of stakeholder-specific knowledge) and
3. Problem (identification/structuring, information needs, underlying assumptions, goal clarity and constraints or resource adequacy).

The framework also considers three main relationships between:

1. Time and social (awareness of potential unintended consequences, sensitivity to changes in stakeholder involvement or perspective over time),
2. Social and problem (socio-political and economic context/issue interaction, perspective-taking, and different framing of problem/goals) and
3. Time and problem (short-long term effects, feedback loop during implementation and self-editing).

Grohs uses a problem scenario to understand undergraduate and graduate students' reasoning processes based on previous assessment tool strategies (e.g., [120]). The hypothetical scenario concerns a problem in a community related to heating, and alluding to social, economic and environmental issues. The measure itself is constructed into three phases:

1. *Processing* to learn how respondents frame the problem, sources of additional information needed, potential measures for success, and groups that should be involved in decision-making
2. *Response* to learn their draft plan for addressing the situation and perceived challenges in implementing the plan
3. *Critique* to learn how respondents analyze and critique an attempted solution to assess the ability to perceive goals of others and identify unintended consequences.

Due to the free-from nature of the responses, the assessment was analyzed through qualitative coding based on the following constructs: problem identification, information needs, stakeholder awareness, goals and measures of success, unintended consequences, and implementation challenges. This assessment was used to qualitatively analyze 27 student responses, and the author argues the data indicates the tool is "an effective means of generating data relevant to specific constructs of systems thinking competency" ([9] p. 74). Furthermore, Grohs suggests that emergent data from each construct indicates patterns exist for structuring the data to study variation across respondents through the establishment of a scoring rubric (which was not discussed in the dissertation).

Systems Thinking Taxonomy

Stave and Hopper [108] offers their taxonomy of a systems thinker, based on their assessment of the literature. This taxonomy, in order of lower order to high order thinking, includes: "recognizing interconnections, identifying feedback, understanding dynamic behavior, and differentiating types of variables and flows, using conceptual models, creating simulation models, and testing policies" [108]. Based on these characteristics, Stave and Hopper [108] provide indicators of achievement for the applicable level and possible products that can serve as an assessment to determine the level of an individual's systems thinking. Stave and Hopper later revised this table based on their analysis of systems thinking interventions in the classroom and their updated understanding of assessment tools to judge systems thinking capabilities [108]. However, Stave and Hopper do not offer an assessment tool to assess these indicators.

Systems Thinking Scale

Moore et. al. developed the Systems Thinking Scale (STS) between 2008 and 2010 to measure the systems thinking capability needed by professionals in the healthcare industry. The STS was built based on the following definition of systems thinking, as derived from an expert panel [121]:

> "The ability to recognize, understand, and synthesize the interactions, and interdependencies in a set of components designed for a specific purpose. This includes the ability to recognize patterns and repetitions in the interactions and an understanding of how actions and components can reinforce or counteract each other. These relationships and patterns occur at different dimensions: temporal, spatial, social, technical or cultural. It is fundamental to undertaking specific

methodology or strategies to explore and redesign a set of components comprising a whole."

The study was also based on "six theoretical dimensions of systems thinking:
1. "Sequence of events
2. Causal sequence
3. Multiple causations possible
4. Variation of different types (random/special)
5. Feedback
6. Interrelations of factors, patterns of relationships" [121].

Moore et. al. [121] report that:

> "the scale is a 20-item instrument using a 5-point likert-type scale. The test-retest reliability assessment (where n=36) had a .74 correlation; internal consistency testing using Cronbach's Alpha (where n=342) had a coefficient of .89. Discriminant validity was tested via a pre/post test with health care professional students (n=102) that received high, low or no dose systems thinking education related to process improvement. The pre-test showed no differences between the STS mean score. The post-test showed that the high dose systems thinking group scored higher than the low or no dose groups. The reliability and validity of the scale were tested based on a 20-item, one-factor instrument. These researchers consulted with eleven experts in the fields of systems thinking and continuous improvement as part of their study. Due to the limited nature of systems thinking and continuous quality improvement measures, concurrent and predictive validity assessments were inconclusive" [121].

Systemic Thinking Inventory for Business

Kurthakoti and Halpin [10] offer a scale to measure systems thinking among business students based on three defined dimensions of systemic thinking (introduced in 2015 as the Systemic Thinking Scale, [11]). Based on an exploratory study, STIB offers three dimensions, and the associated method of analysis for the dimensions:
1. Locus of attention: what a decision maker focuses on while making decisions (i.e., the whole task or its parts).
2. Inter-relatedness: an appreciation of the interconnectedness of the parts of a task or issue.
3. Flexibility: a systemic thinker's creative responsive to changing conditions while solving problems.

Kurthakoti and Halpin [10] report (from an exploratory study with small sample of graduate students) that the final solution resulted in three factors: locus of attention (7 items), inter-relatedness (6 items) and flexibility (3 items), with the factor loadings for individual items ranging between .53 and .84 for both pre- and post-tests.

Psychometric Efforts to Measure Aspects of DSRP

There are examples of psychometric measures that focus on or attempt to measure portions of DSRP Theory. Below we explore some of these assessments.

Distinctions (D) (identity-other)

Our search of the literature yielded no psychometric or edumetric assessments that specifically measure the identity-other elements of Distinction (D) rule. The types of things that would be considered as a skill of this rule include:

1. Distinctions are all around us, it's how we name, identify and differentiate things, ideas, or objects from one another.
2. The identity-other structure of distinctions means that any object or idea is both an identity and an other (e.g., "us" vs. "them").
3. The distinctions you make can be general and/or specific (e.g., "a cup" vs. "a red porcelain cup").
4. Often a single distinction can become many more distinctions when looks closer at its meaning (e.g., "birds" can be further distinguished to be owls, eagles, seagulls).

There is quite a bit of research on people developing their own identity, and how that is developed in relation to others. Psychometric tests are available in this regard, but we did not include them in this lit review, as they do not appear to deal with the cognitive aspects of distinction formation. In a similar vein, in regards to identity formation, research has assessed individual tendencies to make distinctions between the self and the other. Aron and colleagues [122] assessed this distinction-making in the context of close relationships, finding that in these situations individuals often do not distinguish between the self and the other, and instead include the other in the self [122]. In our review of the literature we found no psychometric studies that specifically assess individual skill level in identity-other distinction making.

Systems (S) (part-whole)

Our search of the literature yielded a few psychometric or edumetric assessments that measure *aspects* of the part-whole elements of Systems (S) rule. The types of things that would be considered as a skill of this rule include:

1. Systems are all around us, it's how ideas or objects are organized, grouped or nested with one another.
2. The part-whole structure of systems means that any object or idea is both a part and a whole simultaneously (e.g., a planet is comprised of land and water and is also part of the solar system).
3. In any whole system, you want to identify the relevant parts to better understand that system.
4. The systems rule tells us that we can "zoom in" to see more parts and "zoom out" to see more wholes (e.g., zoom in to see the land and water parts of a planet, zoom out to see that planet as part of the solar system).

Sternberg-Wagner Self-Assessment Inventory

The Sternberg-Wagner Self-Assessment Inventory (SWSAI) (included in [123]), is based on Sternberg's Mental Self-Government (MSG) thinking styles theory [124]. The MSG theory "proposes that the organization of thinking parallels the organization of political government and that individuals govern themselves according to their personal thinking styles or, more accurately, their own profiles of thinking styles" [125]. This theory suggests that the way individuals use their minds is analogous to various dimensions of government. The theory includes a dimension that distinguishes individuals' preference for problems that are global versus those that are local. The inventory is a 104-item psychometric instrument aimed to assess individual's preferences for the thinking styles identified in the MSG theory. The inventory measures each thinking style (e.g., executive, judicial or legislative styles) with eight items, each answered on a Likert scale. Items assessing global thinking styles are relevant to the STMI, as they touch on an individual's ability to organize and understand ideas based on the parts and wholes of the systems. The SWSAI Global thinking style comprises eight items from Sternberg [123]:

1. "I prefer to deal with specific problems rather than general questions.
2. I prefer tasks dealing with a single, concrete problem rather than general or multiple ones.
3. I tend to break down a problem into smaller ones that I can solve, without looking at the problem as a whole.
4. I like to collect detailed or specific information for projects I work on.
5. I like problems where I need to pay attention to detail.
6. I pay more attention to the parts of a task than to its overall significance
7. In discussing or writing on a topic, I think the details and the facts are more important than the overall picture.
8. I like to memorize facts and bits of information without any particular content."

The Inventory has been assessed in a number of studies [126,127]. An early study by Sternberg [128] found the median reliability of the inventory to be .78, with the range between .56 and .88. A study by Dai and Feldhusen [126] determined that evidence only partially supports the internal validity of the instrument.

Analysis-Holism Scale

Choi et al. [129] developed the Analysis-Holism Scale (AHS) to "measure analytic versus holistic thinking tendencies," where individuals engaging in analytic thinking tend to view the world as a set of independent components. The scale is based on the a priori understanding that East Asians think holistically, and thus oriented towards "the relationships between objects and the field to which those objects belong" [129]. While Westerners tend to think analytically, thus focusing "attention more on an object itself rather than on the field to which it belongs" (692). The AHS scale includes four dimensions: "locus of attention (parts vs. whole), causal theory (dispositional vs. interactional), perception of change (linear vs. cyclic), and attitude toward contradictions (formal logic vs. naïve dialecticism)" [129]. The scale was constructed such "individuals with high scores on the scale would give more attention to the whole (rather than parts), explain causal relationships in terms of the interaction between actors and surrounding environments (rather than based on the disposition of actors), exhibit a cyclic (rather than linear) perception of future events, and prefer dialectical (rather than formal) logic in reconciling contradictory propositions" [129]. In this scale, factor 1 (causality) and factor 4 (locus of attention) are related to two of the four cognitive dimensions of the STMI, relationships and systems, respectively. The Cronbach's α for causality and locus of attention are .71 and .56, respectively. The items related to these two factors include from Choi et al. [129]:

"Causality

1. Everything in the universe is somehow related to each other.
2. Nothing is unrelated.
3. Everything in the world is intertwined in a causal relationship.

4. Even a small change in any element of the universe can lead to significant alterations in other elements.
5. Any phenomenon has numerous numbers of causes, although some of the causes are not known.
6. Any phenomenon entails a numerous number of consequences, although some of them may not be known.

Locus of Attention

1. The whole, rather than its parts, should be considered in order to understand a phenomenon.
2. It is more important to pay attention to the whole than its parts.
3. The whole is greater than the sum of its parts.
4. It is more important to pay attention to the whole context rather than the details.
5. It is not possible to understand the parts without considering the whole picture.
6. We should consider the situation a person is faced with, as well as his/her personality, in order to understand one's behavior."

The AHS was tested for convergent and discriminant validity in relation to other identified scales (the Attributional Complexity Scale, the Sternberg-Wagner Self-Assessment Inventory on the Global Style and the Rahim Organizational Conflict Inventory–II). Based on 6 studies performed by Choi et al. [129], the AHS was found to have adequate convergent and discriminant validity, as well as psychometric reliability [129].

Relationships (R) (action-reaction)

Our search of the literature yielded no psychometric or edumetric assessments that specifically measure the action-reaction elements of Relationship (R) rule. The types of things that would be considered as a skill of this rule include:

1. The Relationship rule reminds us to identify and examine the relationships among all the parts of a system. In any system, you want to see not only the nodes - but also the relevant relationships among them to better understand that system.
2. The action-reaction structure of relationships means that any object or idea is an action or reaction (e.g., Person A can act upon Person B or react to Person B).
3. The R rule encourages not only to recognize that a relationship exists but to distinguish that relationship to better understand it (i.e.., by naming it, for example the relationship between "mom" and "dad" is "marriage".)
4. The R rule encourages not only to recognize that a relationship exists but also to zoom into that relationship to see its constituent parts (e.g., the relationship between a farmer and consumer is a vast supply chain made up of many parts; the synaptic relationship between =neurons is made up of electrochemical components).

Perspectives (P) (point-view)

Our search of the literature yielded a few psychometric or edumetric assessments that measure *aspects* of the point-view elements of Perspectives (P) rule. The types of things that would be considered as a skill of this rule include:

1. The Perspectives rule reminds us to examine systems from multiple perspectives to better understand any system.

2. The point-view structure of Perspectives means that any object or idea can be a point and/or a view (e.g., A person (point) can see another person (view); or different states (point) see the parts of marriage (view) differently.)
3. The Perspectives Rule encourages us to take both perspectives "with eyes" (e.g., people, stakeholders, groups, countries, animals) but also non-human perspectives (e.g., economic, political, historical, structural, strengths, weaknesses, color, etc.)
4. When you change the way you look at things (Perspective), the things you look at change (e.g., the Southern perspective on the Civil War includes different things than the Northern perspective on the Civil War).
5. Perspectives can be used as a frame on a system that can either limit/narrow or expand/widen what you see (e.g., looking only at a system from an economic-impact perspective limits what is included while taking a holistic perspective broadens the view).

Researchers have examined perspective-taking as a cognitive process since the early 1900s. Piaget [130] spearheaded this work by examining the role of perspective taking during a child's development. Later, Piaget collaborated with Inhelder to devise the "three mountains task" where children "stood in a spot to view an arrangement of three model mountains and were asked which photograph reflected how the scene would look from the other perspective" [131]. The conclusion of this study found that young children do not have sufficient cognitive development to take another point of view. As they develop, children learn the viewpoints of other people and things [131,132].

Davis [133] suggested that individual difference measures should provide an assessment of the "cognitive, perspective-taking capabilities or tendencies of the individual" separate from the "emotional reactivity of such individuals." Davis explained that "only by separately measuring such characteristics that their individual effects on behavior can be evaluated" [133].

Interpersonal Reactivity Index

The most widely used perspective-taking inventory Davis's [133] Interpersonal Reactivity Index, which includes four seven-item scales, including: perspective-taking; fantasy (tendency to transpose oneself into the feelings and actions of fictitious characters); empathy (other-oriented feelings of sympathy), and; personal distress (feelings of personal anxiety and unease in interpersonal settings). The perspective taking scale, relevant to our analysis, assesses "tendency to spontaneously adopt the psychological point of view of others" [133], pgs. 113-114). The perspective-taking items measure the ability of a respondent to adopt a perspective or a point of view [133]. This scale has been used in a variety of studies to assess perspective-taking (e.g., [134,135]) Davis [133] found "perspective-taking and empathic concern to be significantly and positively related" ($r = .33$) based on two samples (1979, $r = 770$ and 1980, $r = 460$).

The perspective-taking scale items include [133]:

1. Before criticizing somebody, I try to imagine how I would feel if I were in their place.
2. If I'm sure I'm right about something, I don't waste much time listening to other people's arguments. (-)
3. I sometimes try to understand my friends better by imagining how things look from their perspective.
4. I believe that there are two sides to every question and try to look at them both.

5. I sometimes find it difficult to see things from the "other guy's" point of view. (-)
6. I try to look at everybody's side of a disagreement before I make a decision.
7. When I'm upset at someone, I usually try to "put myself in his shoes" for a while.

Previous and subsequent scales have been developed to assess empathy, which while distinguished from perspective-taking, is related in that in a cognitive sense, empathy is an ability to understand one person's perspective. "Perspective taking is a cognitive ability to consider the world from other viewpoints and "allow an individual to anticipate the behavior and reactions of others" ([133] , p. 115). Empathy, in contrast, is an other-focused emotional response that allows one person to affectively connect with another. Sometimes labeled sympathy or compassion, empathy is often considered to be an emotion of concern when witnessing another person's suffering" [134,136,137]. Hogan [138] explains that empathy is, "seen as an everyday manifestation of the disposition to adopt a broad moral perspective, to take 'the moral point of view.'"

Empathy also has an emotional construct, often analyzed separately from its cognitive construct. Scales such as the Hogan Empathy Scale [138] measures the cognitive constructs of empathy. Hogan's scale [138] is based on items developed for the Minnesota Multiphasic Personality Inventory (MMPI) and the California Psychological Inventory [139]. The Mehrabian and Epstein Emotional Empathy Scale [140] measures empathy as an emotional construct (i.e., the tendency to respond emotionally to the experiences of others).

Self Dyadic Perspective-Taking Scale and Other Dyadic Perspective-Taking Scale
Long [141], for example, suggests that perspective taking" may vary across relationships and situations." ([141], pg 93) Long assess dyadic perspective-taking, which "indicates whether or not one individual seeks to understand the point of view of the other person in the dyad. Long developed a Self Dyadic Perspective-Taking Scale (SDPT) provides a self-report that reveals how a person within a relationship perceives the other person's perspective. Long also developed the Other Dyadic Perspective-Taking Scale (ODPT) to assess the extent to which an individual in a dyad is perceived to be a perspective-taker. An example of a SDPT question is "I try to understand my partner better by imagining how things look from his/her perspective". An example of an ODPT question is "When my partner is upset with me he/she tries to put him/herself in my shoes for awhile." The measures were tested with samples of heterosexual dating college students and married individuals. Each scale includes two factors: (1) strategies, which "assess the attempts and endeavors made to understand the point of view of the partner" (pg. 94) and (2) cognizance, which represents "a global understanding and awareness of a partner (pgs. 94-95). For SDPT, on the cognizance subscale the item score correlations ranged from .51 and .72. The alpha coefficient for both partners was .85. On the strategies subscale item score correlations ranged from .47 and .84. The alpha coefficient for both partners was at or above .80. For ODPT, on the cognizance subscale, the item-total score correlation ranged between .53 and .78. The alpha coefficient scores were at or above .87 for both partners. On the strategies subscale, item-total score correlations ranged from .56 and .82. The alpha coefficient was at or above .90 for both partners.

More recent efforts have been made to understand perspective-taking within different contexts. Other studies have sought to understand individual perspective-taking tendencies based on controlled experimental settings, including the role of perspective-taking in increasing contact with individuals typically stereotyped in society [135]. Some of these studies have sought to understand the differences in perspective-taking as a factor of the context or situation (e.g., [131,141,142])

While the above examples provide insight into measurements of *aspects* of DSRP, as a collective they are incomplete and do not represent a purposeful and robust attempt to measure *all* of the variables that underlie systems thinking. The advent of DSRP Theory articulates what these variables are and allows us to measure systems thinking in a comprehensive way.

Introducing the STMI

The Systems Thinking and Metacognition Inventory or 'STMI' [143] is an edumetric test that measures: (1) aptitude in five domains (the four patterns of systems thinking as well as the ability to mix and match the patterns), (2) self-perception of aptitude of five domains, (3) metacognitive (awareness) of these five domains, and (4)[2] others' perception (i.e., 360°) of aptitude and awareness of five domains.

The STMI measures both systems thinking skill, confidence in each skill, and metacognition generally. Because systems thinking is less something we *'do'* than it is an *emergent outcome* of the conscious execution of four cognitive skills, the development of these skills (DSRP) is the primary way to increase the probability of systemic thought. When we consciously employ these cognitive skills we are being metacognitive, or aware of our own thinking processes. Being aware of the distinctions we make, the perspectives we take, the systems of which we are part, and our interrelatedness makes us better systems thinkers.

Because STMI is an edumetric test, it measures growth and the report focuses on areas for potential growth (see Figure 1). The STMI can be taken multiple times, and it is expected that subjects will do better each time, given that the report itself is not only *informative* of results, but also *prescriptive* of future growth and development. The customized report isolates the subject's strengths and weaknesses as well as areas for future improvement and then suggests resources for review. It also compares your actual systems thinking skills in four areas with the subject's self-assessment (confidence) of these same skills. The STMI is an edumetric assessment that measures individual changes in capability. It can be taken multiple times to track one's progress in both understanding and consciously applying systems thinking. The report is split into aggregate and specific scores in 10 dimensions as well as comparisons between dimensional-pairs (skill vs confidence). The report includes assessments of:

1. Understanding of the four cognitive skills underlying systems thinking;
2. Ability to simultaneously apply these skills;
3. Self assessment of your aptitude applying these four skills; and
4. Awareness of one's thinking, or metacognition.

[2] This version of the STMI test has been developed but not yet tested or validated.

RESULTS SUMMARY	May 22nd 2018	May 21st 2018
Aggregate	57/100	89/100
Distinction	73/100	97/100
System	48/100	96/100
Relationship	67/100	96/100
Perspective	50/100	78/100
Mixed	61/100	94/100
Metacognition	33/100	88/100

Figure 1: STMI's edumetric design tracks growth over time

The test is administered entirely online and results are immediately available to the respondent. The STMI has 77 questions: 6 demographic questions; 61 skills questions (D=9, S=10, R=11, P=11, MM=20), and; 10 metacognitive questions. The 360° test is an add-on consisting of 20 questions (7 demographic and 13 other-perception). The STMI typically takes between 15 and 30 minutes to complete. Question types include: true/false; multiple choice (select one) and (select all that apply), and; Likert-scale (Strongly Agree, Agree, Disagree, Strongly Disagree) and (Always True, Mostly True, Sometimes True, Never True). All of the questions have fixed answers (no open ended questions) with the exception of demographic questions. All questions were tested for reading level to a ninth grade literacy level. Figure 2 illustrates one type of skill/aptitude question:

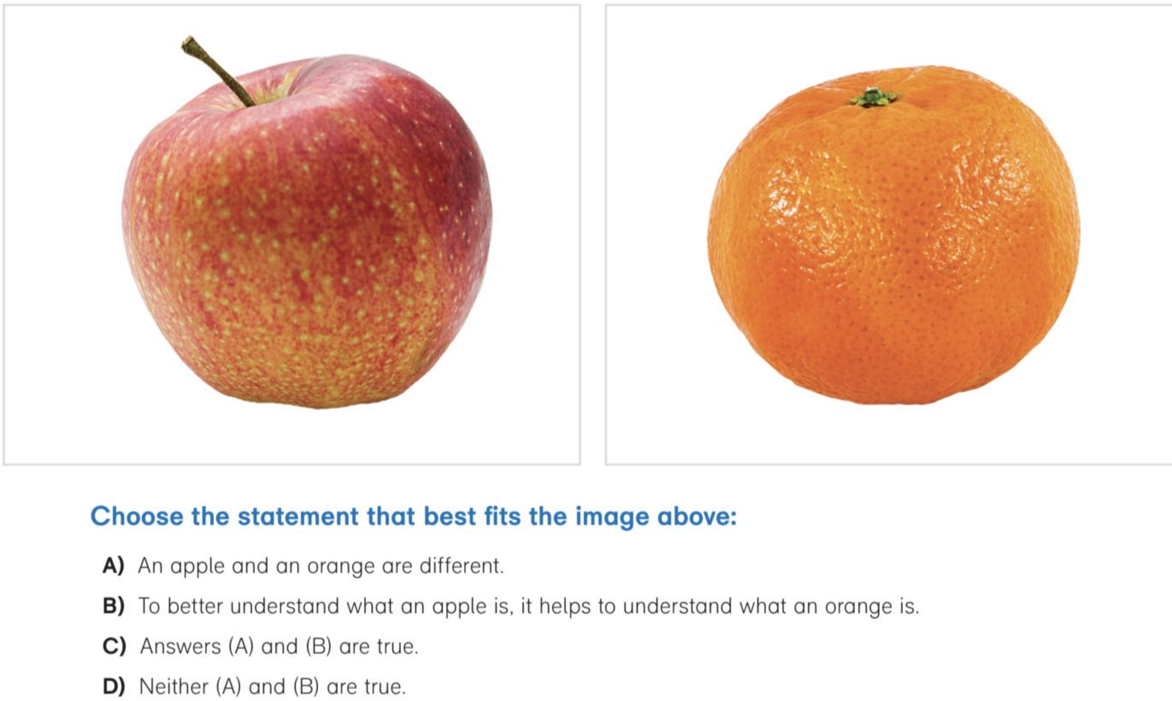

Figure 2: Example multiple choice question from STMI

The questions in the self-perception (and 360° other-perception) section of the test parallel the skill groupings such that the aggregate scoring for the test is delivered as a comparison between skill and confidence across 5 domains.

Scoring Rubric

Assessment scoring can have important implications for validity. Taylor ([144] pg. 99) provides detail about how scoring (grading)[3] rules are important to validity because small changes in answer keys or answer options can change the construct being tested. Ordinal scales can be difficult for assessing score changes over time across groups (e.g., a one unit-change at the top may not equal a one-unit change at the bottom of the range). One should know whether you plan to do criterion versus norm referenced scoring *before* writing the test items "and provide a rationale for the selected scoring model anchored in the stated purpose of the assessment" ([144], pg. 101)[4]. Another scoring issue relates to giving a single score for an exam, which implies unidimensionality of the construct. As a result, five separate scores (D, S, R, P, and Mixed) are calculated for both skill and confidence. This is critical as any given respondent likely has

[3] Norm-referenced tests "allow a comparison of an individual's test performance to the test performance of other individuals. … In other words, did the test taker do better or worse than others who have taken the test?" (Salkind 2012, p. 115) A criterion-referenced test uses cutoff points. For example, it might specify that folks need to get 80% of questions right to demonstrate proficiency.

[4] Some of the types of evidence Taylor ([144], pg. 111) reports as necessary to establish the validity of a test are: "statistical item-analyses of data; how item data are used to evaluate the quality of items, scoring rubrics, rating scales, and distractors for multiple-choice items and to select items for the final form of the assessment; rationales for scaling and scoring models; evidence that scoring rules can be applied consistently across scorers."

strengths in some of the skills and is seeking diagnosis of areas of growth to become a better systems thinker and more metacognitive. As such, separate items measure metacognition using *modified* items from the previously validated MAI (See above or [67] pg. 472).

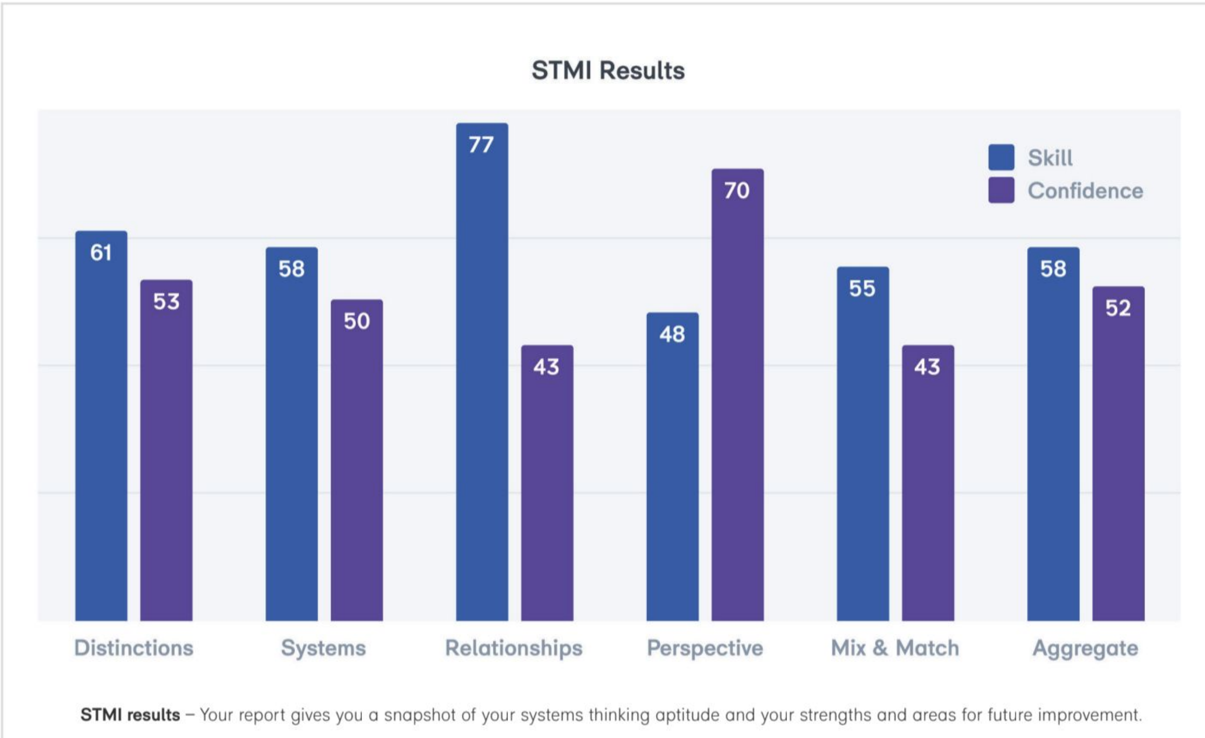

Figure 3: Comparative Skill and Confidence scores (10 dimensions) and aggregate scores

While the aggregate skill score can be useful (particularly tracked over time), keep in mind that this single number can mask important differences between the individual scores on each skill. Thus, subjects are encouraged to analyze their individual skills and confidence scores across D, S, R, P and Mix and Match dimensions (Figure 3).

When respondents compare skills and confidence scores they are encouraged to look for areas where the difference between the scores is high. A large difference in the scores indicates a 'blind spot.' For example, if your perspective skill score is relatively low (say, 60) and your perspective confidence score is also relatively low (say, 62) this shows that you are estimating your skills relatively accurately. Whereas if your skills score is low and your confidence score is high, this means you are *overestimating* your skill in that area. Likewise, if your confidence score is relatively low compared to your skill score this means you are *underestimating* your skill in that area. Metacognitive awareness (and accuracy of awareness) is, in many ways, as relevant to growth and development as are raw skills scores. So any blind spot—both overestimating or underestimating skill—is an area of potential growth, as is the development of the skill itself.

The STMI test is designed to increase the test takers' understanding of skill gaps and receive a "pros and grows analysis." Rather than an MBTI indicator of your systems thinking 'type' it is providing a snapshot in time of your aptitude for each skill of systems thinking, confidence in your abilities with each skill, and metacognitive awareness of each skill. The STMI can be administered before and after exposure to systems thinking to see one's improvement. Immediate results are provided in the form of an online report. The report gives a snapshot of the test taker's systems thinking skills and confidence, weaknesses and strengths, and areas for future improvement. In this way, the report is a form of treatment. For individuals and organizations, the purpose of the test is to establish baselines and motivate improvement. The Systems Thinking Metacognitive Inventory (STMI) is designed to measure your strengths and weaknesses in the metacognitive skills that underlie systems thinking. It provides a 'pros and prows' analysis of where you're strong and where you can improve on 10 dimensions.

Validation process

Creation of the STMI involved the following steps: development of theoretical and conceptual underpinnings (2002-2012), articulation of underlying theory (DSRP, 2005), translation of concepts into practice (2005-2012) in elementary schools, translate concepts into measurable skills (2013-2014), preliminary questionnaire development (2014) and score coding, pilot testing (2014-2016) was conducted for feedback and comments about any issues/concerns they have with any of the survey questions -- prior to programming the instrument for the larger web-based pilot test, this led to instrument revision, focus group testing (August 2016), preliminary tests of validity (2016-2017 sample=100), final review, instrument (added new items explicitly to capture metacognition based on extensive literature review of other measures and using MAI items to increase criterion validity) and score coding revision (2018), and then additional testing with samples of 1000 (2019).

Questionnaire Development and Initial Testing

Questionnaire Development: Developed an inventory of items in the form of a questionnaire to measure skill in making distinctions (D), organizing systems (S), recognizing relationships (R), and taking perspectives (P) and their combined (integrative) use. Additional items measure respondent's confidence in each of these four skills and their integration that underlie systems thinking. After a few rounds of initial testing, additional items were included to garner self-reported data to capture the metacognition levels of individual respondents. Thus, the instrument included specific items to measure each of the 10 dimensions (D, S, R, P and I, each having two levels, skill and confidence) as well as metacognitive questions based on the MAI. Note as well, that during this development phase, emphasis was placed on designing items that had both a visual and linguistic components to increase the internal construct validity and reduce respondent biases.

STMI v2.0 early testing: Data gleaned from STMI v. 1.0 allowed for the next draft of the instrument to be whittled down to 74 questions, and reduce the completion time to approximately 40 minutes. Primary changes made included removing the questions that contribute least to the validity of constructs based on statistical outputs. Thus, the new version was hypothesized to increase validity while also reducing size and difficulty of the instrument, In late 2016, The survey was then input to Survey Monkey for ease of testing, data collections and analysis. The test was 74 questions in length, with 5 concerning demographics, 55 skill based questions on

D,S, R, P and mixing and matching; and 14 items that were self-report of conscious application of and aptitude with DSRP (at this time called "metacognition" specific to each pattern).

STMI v2.1: also used Survey Monkey and was administered (early 2017) to a randomly sampled population of 99 adults. Respondents were allowed to skip questions, and the number of responses for each question ranged from 77 to 89. Version 2.1 was further reduced to 64 questions in hope of getting response time down to 30 minutes. Approximately 47 substantive questions were analyzed to test the veracity of our D, S, R, and P scales. The changes made since the survey was previously administered significantly raised Cronbach's for systems but lowered it for relationships. The analysis produced the following alpha coefficients for DSRP (the number of questions ("items") and cases are in parentheses):

- Distinctions: .48 (7 items; 90 cases)
- Systems: .63 (10 items; 89 cases)
- Relationships: .28 (8 items; 82 cases)
- Perspectives: .52 (10 items; 77 cases)

The changes made between versions 2.0 and 2.1 to the relationships index significantly lowered its alpha. We calculated that returning 3 true/false relationships rule questions to the survey would increase the alpha for the R index to .566 or greater.

At this time, STMI v2.1 was administered to several test audiences from various subgroups including graduate students, academic fellows, practitioners, educators, government professionals and corporate managers for an additional 246 respondents. This data was used to further test the utility and ease of administering the test, as well as to create a scoring rubric, and to generate an immediate accessible report for each individual respondent.

Scoring Rubric & Coding: Of equal complexity was the development of a robust scoring rubric to be coded based on allocations of quantitative scores for each possible response to each item. Notably, many of the skills measured by the STMI exist on a continuum, meaning, a respondent exhibits strengths in some and possible weaknesses in others. As a result, the scoring rubric reflected this contimmu in that often a particular answer was "more correct" than another and thus received a higher quantitative score to reflect that - while other responses received a lower quantitative score but none of the items (other than True/False) were scored as binary in the assigned score. This score coding scheme allowed for a greater understanding of the differences among respondents; but also for *each respondent among the skills tested.* Thus, a respondent received a diagnostic score that reflected areas of competence (DSRP & I) and also areas for growth.

Focus Group (August 2016): As the next step in this initial questionnaire development phase, 20 adult respondents were recruited to participate in an in-person focus group session. The purpose of the focus group session was to cognitively test the instrument items to ensure individuals understood the meaning of each item as intended. Participants were asked to complete the questionnaire and comment on each item about which they were unclear. At the end of the focus group session, a summary of those items that were unclear or needed revision was established.

STMI v. 2.2 Web Pilot Test (Dec 2016): A more formalized pilot test of this revised questionnaire was conducted with a sample of 100 completed surveys. STMI V2.2 incorporated the suggested revisions from 2.1 and was administered to a web panel of 100 adults. At this time, the instrument was programmed for web administration. A web panel of 100 adult respondents was purchased for this pilot test. At the end of data collection, summary statistics on each item were generated, and inter- and intra-item correlations were computed. Additional statistical analyses identified those items that contributed most to the variance in the responses. As a result, several items were revised and/or removed from the instrument.
Final Review/Revision: Combining the web pilot test results and the focus group results allowed for significant revision to the STMI instrument that necessitated corollary revisions to the scoring system.

STMI v. 3.0 Web Pilot Test (2019): A web panel of adults ages 18-85 in the continental U.S. was purchased from TurkPrime, providing a viable sample of 1000 respondents. The panel was a non-probability based sample, and the goal was to use the data for development purposes only. This required the instrument to average no more than 30 minutes to complete.

Table 2 below shows the corresponding sample sizes, Cronbach's Alpha Coefficients and related items in each version of the STMI. Note that STMI v.2.2 resulted in an overall xxxxx reliability in *r*.66 as indicated by Cronbach's alpha coefficient which is consistent with similar scores in validated tests mentioned in Liu et. al. [61]. Of the 57 questions that make up the skill section of the test, these are grouped into 5 groups: D, S, R, P, and M. The coefficients for these groupings in each version of the instrument are as follows:

STMI Version	**Sample Size**	**Cronbach's Alpha Coefficient**	**Number of items**
v1.0	n=76	Distinctions: .46 Systems: .19 Relationships: .57 Perspectives: .64	(D) 17 items (S) 15 items (R) 14 items (P) 21 items
v2.1	n=99	Distinctions: .48 Systems: .63 Relationships: .28 Perspectives: .52	(D) 7 items (S) 10 items (R) 8 items (P) 10 items
v2.2	n=100	Distinctions: .51 Systems: .38 Relationships: .56 Perspectives: .64 Mix & Match: .46 Whole Instrument: .66	(D) 17 items (S) 13 items (R) 15 items (P) 19 items (MM) 20 items
v3.0	n=1059	Distinctions: .54 Systems: .61 Relationships: .58	(D) 9 items (S) 10 items (R) 11 items

		Perspectives: .64 Mix & Match: .50 Whole Instrument: .69	(P) 11 items (MM) 20 items (M) 10 items

Table 2: Sample size, Cronbach's Coefficients & Items per STMI version

Population Characteristics: Systems Thinking

Skill and confidence comparisons across (n=1059) respondents are shown in Figure 4[5]. The confidence scores (depicted as light gray bars) are significantly and consistently higher than the skills scores (dark gray bars). One is reminded of the Dunning-Kruger effect [145], in which a "the miscalibration of the incompetent stems from an error about the self, whereas the miscalibration of the highly competent stems from an error about others." It appears that on average, in systems thinking, people have more confidence than skill across all domains.

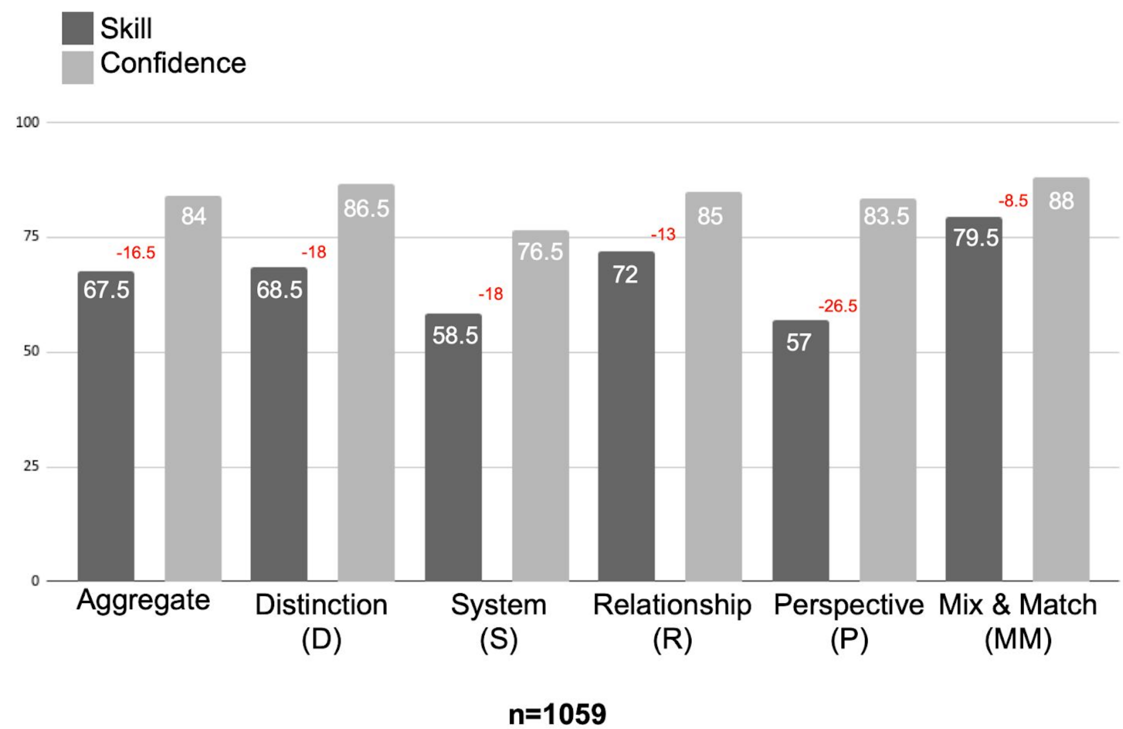

Figure 4: General Skill and Confidence Comparisons Across (n=1059)

Figure 4 also provides evidence that across the sample (n=1059) a general population of respondents (none having formal training in systems thinking) demonstrate a baseline level of skill in DSRP. This is congruent with evidence that DSRP exists as universal patterns of mind

[5] Skill and confidence comparisons across (n=1059) respondents are shown in Figure 4 and numerically as follows: Skill-aggregate=67.5, confidence-aggregate=84, difference=-16.5; skill-distinction=68.5, confidence-distinction=86.5, difference=-18; skill-system=58.5, confidence-system=76.5, difference=-18; skill-relationship=72, confidence-relationship=85, difference=-13; skill-perspective=57, confidence-perspective=83.5, difference=-26.5; skill-mixed=79.5, confidence-mixed=88, difference=-8.5.

[88]. The relative lack of skill shown in Table 4 also shows that, while all people are doing a substantial amount of DSRP thinking as a baseline, they can also improve it, by increasing their metacognitive awareness of the DSRP patterns.

Next Steps/Further Work

Progress on the Systems Thinking and Metacognition Inventory (STMI) is steady, but there are a few issues that must be addressed in future work. For example, it is well-understood that self-report questions can be unreliable due to the respondent's desire to be perceived in a positive light and therefore to select the 'best' answer rather than the 'true' answer. This issue does not apply to the skill portion of the STMI test, but it does apply to the self-perception or confidence section of the test. We also hope to continue to increase both the reliability and validity of the test. Over several versions and related testing of various samples (totalling 1,580 individuals over time) we have been able to significantly raise the reliability to a respectable level (.69), so we are confident that the test will increase in its reliability and validity with further work. As part of this validation, we have yet to validate the STMI in relation to other validated tests in metacognition, critical thinking, IQ, or tests like the SAT and GRE.

We have also focused on making the test not only scientifically valid and reliable but also 'fun to take.' Although the measurement of 'fun' is not typically considered in an academic setting, we think that it is a crucial part of any test, as it is likely that test fatigue and boredom play a role in the reliability and authenticity of respondents' answers. Making the process engaging, fun, and interesting is important—but it is difficult, when trying to simultaneously ensure validity and reliability. While much of this 'fun factor' has to do with design work on individual questions and interface, it also has to do with the results report itself . We believe that—especially in regard to an edumetric test—the report is as important as the test because it provides a diagnostic for each respondent and offers feedback on ways to improve systems thinking skills. The STMI report provides the test taker with an assessment of their skill and confidence, and it empowers them to have personal agency in changing or improving that assessment. Thus, the report is the first post-test treatment as well as a treatment for an ensuing retake test. In short, the test-report combination should *facilitate* and *motivate* the overall improvement of systems thinking. 'Fun' (or engaging), in the current culture of multitasking and snippet video, can also mean 'short.' It is also a challenging task to shorten tests while maintaining their reliability, validity, and scope. Increases in fun and decreases in length and time-on-task mean larger sample sizes, more data, and more systems thinkers.

There are additional dimensions to test that are not captured explicitly in the current test. Specifically, the dimension of ethics that arises out of systems thinking-DSRP [72]. This is an area that is sorely needed in society writ large and we would hope to include this dimension in future tests. Additionally, we plan to test the untested 360° other-perception of the test—this will allow for social feedback for individuals and in organizations and teams. As systems thinking gains popularity and promise, identifying its foundational elements offered by DSRP requires corresponding methods of both teaching and assessing systems thinking as a much needed skill.

References

1. Cabrera D, Cabrera L. Systems Thinking Made Simple: New Hope for Solving Wicked Problems in a Complex World. Odyssean Press. Ithaca: Odyssean Press; 2015.

2. Cabrera D, Milstein B, Gallagher RS, Leischow S. Practical Challenges of Systems Thinking and Modeling in Public Health. American Journal of Public Health. 2006;96: 1–9.

3. Cabrera D. What is Systems Thinking? Cornell University. 2005.

4. Cabrera D, Cabrera L, Lobdell C. Systems Thinking. Journal of Evaluation and Program Planning. 2008;31: 299–310.

5. Cabrera D, Cabrera L. Distinctions, Systems, Relationships, and Perspectives (DSRP): A Theory of Thinking and of Things. Journal of Evaluation Program Planning. 2008;31: 311–17.

6. Cabrera D, Cabrera L. Thinking at Every Desk: Four Simple Skills to Transform Your Classroom. Odyssean Press; 2008.

7. Betts F. How Systems Thinking Applies to Education. Improving School Quality. 1992;50: 38–41.

8. Lannon C. Revitalizing the Schools: A Systems Thinking Approach - The Systems Thinker. In: The Systems Thinker [Internet].

9. Grohs JR, Doolittle-Chair PE, Fowler SB, Jones BD, Magliaro SG. Developing a Measure of Systems Thinking Competency. In Education. 2015.

10. Kurthakoti R, Halpin AL. STIB : The Systemic Thinking Inventory For Business. 2016;43.

11. Halpin AL, Kurthakoti R. Can Systemic Thinking Be Measured? Introducing The Systemic Thinking Scale (STS). Developments in Business Simulation and Experiential Learning. 2015;42: 151–153.

12. Carver R. Two Dimensions of Tests. 1974.

13. Borsboom D, Molenaar D. Psychometrics. International Encyclopedia of the Social & Behavioral Sciences. 2015;18: 418–422.

14. Darwin C, Mayr E. On the origin of species : A facsimile of the first edition. Cambridge: Harvard University Press; 1964. pp. xxvii, ix, 502 , [1] folded leaf of plates.

15. Waller JCW. Galton, Sir Francis (1822–1911). International Encyclopedia of the Social & Behavioral Sciences. 2015;9: 569–572.

16. McCaulley MH. The Myers-Briggs Type Indicator and leadership. In: Clark KE, editor. Measures of leadership , (pp. 1990. pp. 381–418.

17. Nowack K. Is the Myers Briggs Type Indicator the Right Tool to Use? Performance in Practice, American Society of Training and Development. 1996;Fall 1996.

18. Psychological Testing: Myers-Briggs Type Indicator.

19. Sternberg RJ, Jarvin L. Binet, Alfred (1857–1911). International Encyclopedia of the Social & Behavioral Sciences. 2015;2: 596–599.

20. Cattell RB. Theory of fluid and crystallized intelligence: A critical experiment. J Educ Psychol. 1963;54: 1–22.

21. Cattell RB. A systems theory of maturation and structured learning. New York: Springer Pub. Co.; 1980. p. xxv, 643 p.

22. Dickens WT. Cognitive Ability. In: Durlauf S, editor. The New Palgrave Dictionary of Economics.

23. Fancher RE, Rutherford A. Pioneers of Psychology: A History. W.W. Norton; 2012.

24. Stern E. Intelligence, Prior Knowledge, and Learning. In: Wright JD, editor. International Encyclopedia of the Social & Behavioral Sciences 2nd edition. Oxford: Elsevier; 2015. pp. 323–328.

25. Kaufman AS, Lichtenberger EO. Assessing Adolescent and Adult Intelligence. John Wiley & Sons; 2005.

26. Braden JP, Niebling BC. Using the joint test standards to evaluate the validity evidence for intelligence tests. In: Flanagan DP, Harrison PL, editors. Contemporary Intellectual Assessment: Theories, Tests and Issues (2nd ed). Guilford; 2005. pp. 615–630.

27. Borsboom D, Mellenbergh GJ, Van Heerden J. Different Kinds of DIF: A Distinction Between Absolute and Relative Forms of Measurement Invariance and Bias.

28. Yen C-J, Konold TR, McDermott PA. Does learning behavior augment cognitive ability as an indicator of academic achievement? J Sch Psychol. 2004;42: 157–169.

29. Elliott CD. The Nature and Structure of Children’s Abilities: Evidence from the Differential Ability Scales. J Psychoeduc Assess. 1990;8: 376–390.

30. Schmidt FL. The Role of General Cognitive Ability and Job Performance: Why There Cannot Be a Debate. Hum Perform. 2002;15: 187–210.

31. Bailin S. Critical and Creative Thinking. 1987;1.

32. Zeng L, Proctor RW, Salvendy G. Can Traditional Divergent Thinking Tests Be Trusted in Measuring and Predicting Real-World Creativity? 2011.

33. Hocevar D, Bachelor P. A Taxonomy and Critique of Measurements Used in the Study of Creativity. Handbook of Creativity. 1989. pp. 53–75.

34. Hocevar D. A Comparison of Statistical Infrequency and Subjective Judgment as Criteria in the Measurement of Originality. Journal of Personality Assessment. 1979. pp. 297–299.

35. Hocevar D. Measurement of creativity: review and critique. J Pers Assess. 1981;45: 450–464.

36. Hughes DJ, Furnham A, Batey M. The structure and personality predictors of self-rated creativity. Thinking Skills and Creativity. 2013;9: 76–84.

37. Batey M, Furnham A. Creativity, Intelligence, and Personality: Critical Review of the Scattered Literature. Genet Soc Gen Psychol Monogr. 2006;132: 355–429.

38. Batey M, Chamorro-Premuzic T, Furnham A. Individual Differences in Ideational Behavior: Can the Big Five and Psychometric Intelligence Predict Creativity Scores?

39. Batey M. The Measurement of Creativity: From Definitional Consensus to the Introduction of a New Heuristic Framework. 2012.

40. Silva LS, Barreto SM. Stressful working conditions and poor self-rated health among financial services employees. Rev Saude Publica. 2012;46: 407–416.

41. Barakat S, Boddington M, Vyakarnam S. Measuring entrepreneurial self-efficacy to understand the impact of creative activities for learning innovation. International Journal of Management Education. 2014;12: 456–468.

42. Runco MA, Acar S. Divergent Thinking as an Indicator of Creative Potential. Creat Res J. 2012;24: 66–75.

43. Weisberg R. The study of creativity: From genius to cognitive science. International Journal of Cultural Policy. 2010;16: 235–253.

44. Plucker JA, Makel MC. Chapter 3: Assessment of Creativity.

45. Guilford JP. The nature of human intelligence. 1967.

46. Runco MA, Millar G, Acar S, Cramond B. Torrance tests of creative thinking as predictors of personal and public achievement: A fifty-year follow-up. Creat Res J. 2010;22: 361–368.

47. Kim KH, Cramond B, Bandalos DL. The Latent Structure and Measurement Invariance of Scores on the Torrance Tests of Creative Thinking–Figural.

48. Treffinger DJ, Young GC, Selby EC, Shepardson C. Assessing creativity: A guide for educators. Storrs, CT: National Research Center on the Gifted and Talented; 2002 p. 97.

49. Torrance EP. Scientific Views of Creativity and Factors Affecting Its Growth. Daedalus. 1965;94: 663–681.

50. Torrance EP, Paul Torrance E. Retooling Education for Creative Talent: How Goes It? Gifted Child Quarterly. 1974. pp. 233–239.

51. Isaksen SG, Treffinger DJ. Creative problem solving. The Basic Course New York: Bearly. 1985.

52. Mednick S. The associative basis of the creative process. Psychological Review. 1962. pp. 220–232.

53. Wallach MA, Kogan N. Modes of thinking in young children. 1965.

54. Ayas MB, Sak U. Objective measure of scientific creativity: Psychometric validity of the Creative Scientific Ability Test. Thinking Skills and Creativity. 2014;13: 195–205.

55. Stupple EJN, Maratos FA, Elander J, Hunt TE, Cheung KYF, Aubeeluck AV. Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking. Thinking Skills and Creativity. 2017;23: 91–100.

56. Facione PA. The California Critical Thinking Skills Test --0211ege Level Factors Predictive of CT Skills The California Critical Thinking Skills Test: College Level Technical Report #2 Factors Predictive of CT Skills.

57. Bonnefon JF. The pros and cons of identifying critical thinking with system 2 processing. Topoi.

2018.

58. Halpern DF. Thought and Knowledge: An Introduction to Critical Thinking. 5th ed. Psychology Press; 2013.

59. Evans JSBT, St. B. T. Evans J, Stanovich KE. Dual-Process Theories of Higher Cognition. Perspectives on Psychological Science. 2013. pp. 223–241.

60. Stupple EJN, Ball LJ. The intersection between Descriptivism and Meliorism in reasoning research: further proposals in support of “soft normativism.” Front Psychol. 2014;5: 1269.

61. Liu OL, Frankel L, Roohr KC. Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment. ETS Research Report Series. 2014;2014: 1–23.

62. Verburgh A, François S, Elen J, Janssen R. The Assessment of Critical Thinking Critically Assessed in Higher Education: A Validation Study of the CCTT and the HCTA. Education Research International. 2013;2013.

63. Ennis RH. Critical Thinking Assessment. Theory Pract. 1993;32: 179–186.

64. Schraw G, Moshman D. Metacognitive Theories.

65. Young A, Fry JD. Metacognitive awareness and academic achievement in college students. Journal of the Scholarship of Teaching and Learning. 2008;8: 1–10.

66. Dwyer CP, Hogan MJ, Stewart I. An integrated critical thinking framework for the 21st century. Thinking Skills and Creativity. 2014;12: 43–52.

67. Schraw G, Dennison RS. Assessing metacognitive awareness. Contemporary Educational Psychology. 1994. pp. 460–475.

68. Oguz Akturk A, Sahin I. Literature Review on Metacognition and its Measurement. Procedia Soc Behav Sci. 2011;15: 3731–3736.

69. Schraw G. A conceptual analysis of five measures of metacognitive monitoring. Dunn Keren Nelson Prins et al. 2005.

70. Baker L, Cerro LC. Assessing Metacognition in Children and Adults. In: Gregory Schraw, James C. Impara, editors. Issues in the Measurement of Metacognition. Lincoln, NE: Buros Institute of Mental Measurements; 2000.

71. Sperling RA, Howard BC, Staley R, DuBois N. Metacognition and Self-Regulated Learning Constructs. Educ Res Eval. 2004;10: 117–139.

72. Cabrera D, Cabrera L, Powers E. A Unifying Theory of Systems Thinking with Psychosocial Applications. Syst Res Behav Sci. 2015;32.

73. Fernández-Berrocal P, Checa P. Editorial: Emotional Intelligence and Cognitive Abilities. Front Psychol. 2016;7: 955.

74. Pérez JC, Petrides KV, Furnham A. Measuring Trait Emotional Intelligence. In: Schulze R, Roberts RD, editors. Emotional intelligence: An international handbook. Ashland, OH: Hogrefe & Huber; 2005. pp. 181–201.

75. Brackett M, Delaney S, Salovey P. Emotional Intelligence.

76. Davies M, Stankov L, Roberts RD. Emotional intelligence: in search of an elusive construct. J Pers Soc Psychol. 1998;75: 989–1015.

77. Matthews G, Zeidner M, Roberts RD. Measuring emotional intelligence: Promises, pitfalls, solutions. Handbook of methods in. 2007.

78. Mayer J, Caruso D, Salovey P. Emotional Intelligence Meets Traditional Standards for Intelligence. Intelligence. 1999;27: 267–298.

79. Mayer JD. MSCEIT: Mayer-Salovey-Caruso emotional intelligence test. Toronto, Canada: Multi-Health Systems. 2002.

80. Brackett MA, Mayer JD, Warner RM. Emotional intelligence and its relation to everyday behaviour. Pers Individ Dif. 2004;36: 1387–1402.

81. Brackett MA, Mayer JD. Convergent, Discriminant, and Incremental Validity of Competing Measures of Emotional Intelligence. Personality and Social Psychology Bulletin. 2003. pp. 1147–1158.

82. Cabrera D. Remedial genius: think and learn like a genius with the five principles of knowledge. Loveland, CO: Project N Press; 2000.

83. Cabrera D. Systems Thinking. Cornell University; 2006.

84. Cabrera D, Colosi L, Lobdell C. Systems thinking. Eval Program Plann. 2008;31: 299–310.

85. Cabrera, D., Cabrera, L. Any Person, Any Study: A Different Kind of Theory of Everything (ToE). In: Cabrera D, Cabrera L, Midgley G, editors. Routledge Handbook of Systems Thinking. Routledge; 2021.

86. Cabrera D, Cabrera L, Midgley G. The Four Waves of Systems Thinking. In: Cabrera D, Cabrera L, Midgley G, editors. The Routledge Handbook for Systems Thinking. London, UK: Routledge Press; 2021.

87. Cabrera D, Cabrera L, Midgley G, editors. Routledge Handbook of Systems Thinking. London, UK: Routledge Press;

88. Cabrera D, Cabrera L, Cabrera E. A Literature Review of the Universal Patterns and Atomic Elements of Complex Cognition. In: Cabrera D, Cabrera L, Midgley G, editors. Routledge Handbook of Systems Thinking. Routledge; 2021.

89. Cabrera D, Arnold L, Cabrera L. The Simple Rules of Complex Networks: Heuristics for Structural Predictions. In: Cabrera D, Cabrera L, Midgley G, editors. Routledge Handbook of Systems Thinking. Routledge; 2021.

90. Cabrera D, Cabrera L. Structural Predictions: Part One. Systems Thinking Daily. 2020; 4.

91. Cabrera D, Cabrera L. Synapses, Hinges, Embassies, and Supply Chains: The Power of RDS Structural Predictions. Systems Thinking Daily. 2020; 4.

92. Cabrera D, Cabrera L. Structural Predictions: Part Two. Systems Thinking Daily. 2020; 4.

93. Boulding KE. General Systems Theory—The Skeleton of Science. Manage Sci. 1956;2: 197–208.

94. Cabrera D, Cabrera L. Thinking at Every Desk: Four Simple Skills to Transform Your Classroom (Second Edition). Norton Books; 2011.

95. Cabrera D, Cabrera L, Kushner J, Hoard D, Solin J, DiBiase M. Systems Thinking: A Little Film About a Big Idea. Photosynthesis Productions; 2015.

96. Cabrera, D. Thought leaders and educators gather to prescribe new flow for water education. Systems Thinking Daily. 2014;14: 10.

97. ___. Cabrera Keynotes AECT Showcasing New Technologies and Wins Outstanding Book Award for Systems Thinking Approach to Education. Systems Thinking Daily. 2017;17: 11.

98. Powers E, Kushner J, Solin J, Cabrera L, Cabrera D. Applying Systems Thinking to Wicked Water Problems: An Accelerated Introduction for Researchers. Journal of Contemporary Water Research and Education. 2018.

99. Cabrera D, Cabrera L. Complexity and Systems Thinking Applications in Education. In: M. S, B. L, M C, editors. Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy. Springer; 2019.

100. Cabrera D, Cabrera L. What Is Systems Thinking? Learning, Design, and Technology. Springer; 2019.

101. Cabrera D, Cabrera L. Complexity and Systems Thinking Models in Education: Applications for Leaders. 2019.

102. Cabrera D, Cabrera L, Kushner J, Hoard D, Solin J, DiBiase M. Count Me In. Cabrera Research Lab; 2014.

103. Cabrera D, Cabrera L. Systems Thinking and Metacognitive Inventory (STMI): A Validated, Edumetric Test of Systems Thinking and Metacognitive Skills. Cabrera Research Lab; 2013.

104. Cabrera D, Cabrera L. Personal Mastery of Systems Thinking in the 4th Wave. In: Cabrera D, Cabrera L, Midgley G, editors. Routledge Handbook of Systems Thinking. Routledge; 2021.

105. Cabrera D, Cabrera L. Complexity and Systems Thinking Models in Education: Applications for Leaders. In: Spector MJ, Lockee BB, Childress MD, editors. Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy. Cham: Springer International Publishing; 2019. pp. 1–29.

106. Cabrera L, Cabrera D. Complex Adaptive Systems (CAS) & Agent Based Modeling (ABM): The Micro Begets the Macro. In: Cabrera L, Cabrera D, Midgley G, editors. The Routledge Handbook for Systems Thinking. London, UK: Routledge Press;

107. Doyle JK. The cognitive psychology of systems thinking. System Dynamics Review. 1997;13: 253–265.

108. Hopper M, Stave KA. Assessing the Effectiveness of Systems Thinking Interventions in the Classroom.

109. Hopper M, Stave KA. Assessing the effectiveness of systems thinking interventions in the

classroom. 26th international conference of the system dynamics society. 2008.

110. Maani KE, Maharaj V. Links between systems thinking and complex decision making. System Dynamics Review. 2004;20: 21–48.

111. Huz S, Andersen DF, Richardson GP, Boothroyd R. A framework for evaluating systems thinking interventions: An experimental approach to mental health system change. System Dynamics Review. 1997;13: 149–169.

112. Cavaleri S, Sterman JD. Towards evaluation of systems thinking interventions: A case study. System Dynamics Review. 1997;13: 171–186.

113. Dorani K, Mortazavi A, Dehdarian A, Mahmoudi H, Khandan M, Mashayekhi AN. Developing Question Sets to Assess Systems Thinking Skills.

114. Richmond B. Systems Thinking - Critical Thinking Skills For The 1990s And Beyond. System Dynamics Review. 1993;9: 113–133.

115. Sterman J. Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill/Irwin; 2000.

116. Senge PM. The Fifth Discipline Fieldbook: Strategies and Tools for Building a Learning Organization. Crown; 2014. J

117. Plate R. Assessing individuals' understanding of nonlinear causal structures in complex systems. Syst Dyn Rev. 2010;26: 19–33.

118. Hoz R, Bowman D, Kozminsky E, Hoz1 R, Bowman2 D, Kozminsky1 E. The differential effects of prior knowledge on learning: A study of two consecutive courses in earth sciences The differential effects of prior knowledge on learning: A stud two consecutive courses in earth sciences. Source: Instructional Science Instructional Science. 2001;29: 187–211.

119. Greene BA, Lubin IA, Slater JL, Walden SE. Mapping Changes in Science Teachers' Content Knowledge: Concept Maps and Authentic Mapping Changes in Science Teachers' Content Knowledge: Concept Maps and Authentic Professional Development. Source: Journal of Science Education and Technology J Sci Educ Techno]. 2013;22.

120. Shin N, Jonassen DH, McGee S. Predictors of well-structured and ill-structured problem solving in an astronomy simulation. J Res Sci Teach. 2003;40: 6–33.

121. Moore SM, Komton V, Adegbite-Adeniyi C, Dolansky MA, Hardin HK, Borawski EA. Development of the Systems Thinking Scale for Adolescent Behavior Change. West J Nurs Res. 2018;40: 375–387.

122. Aron A, Aron EN, Tudor M, Nelson G. Close Relationships as Including Other in the Self. J Pers Soc Psychol. 1991;60: 241–253.

123. Sternberg RJ. Thinking Styles. 1997.

124. Sternberg RJ. Mental Self-Government: A Theory of Intellectual Styles and Their Development. Hum Dev. 1988;31: 197–224.

125. Black AC, Mccoach DB. Validity Study of the Thinking Styles Inventory. Journal for the

Education of the Gifted. 2008;32: 180–210.

126. Dai DY, Feldhusen JF. A validation study of the thinking styles inventory: Implications for gifted education. Roeper Rev. 1999;21: 302–307.

127. Kaufman JC. Narrative and Paradigmatic Thinking Styles in Creative Writing and Journalism Students. J Creat Behav. 2002;36: 201–219.

128. Sternberg RJ. … styles: Theory and assessment at the interface between intelligence and personality. W: RJ Sternberg, P. Ruzigs (red.), Intelligence and personality. 1994.

129. Choi I, Koo M, Jong An Choi. Individual differences in analytic versus holistic thinking. Pers Soc Psychol Bull. 2007;33: 691–705.

130. Piaget J. The Moral Judgment of the Child. Free Press; 1965.

131. Parker SK, Axtell CM. Seeing another viewpoint: Antecedents and outcomes of employee perspective taking. Acad Manage J. 2001;44: 1085–1100.

132. Piaget J, Inhelder B, Sinclair-de Zwart H. Mémoire et intelligence. 1968.

133. Davis MH. Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. J Pers Soc Psychol. 1983;44: 113–126.

134. Galinsky AD, Maddux WW, Gilin D, White JB, White4 JB. Why It Pays to Get Inside the Head of Your Opponent in Negotiations. Psychol Sci. 2008;19: 378–384.

135. Wang CS, Tai K, Ku G, Galinsky AD, Urgesi C. Perspective-Taking Increases Willingness to Engage in Intergroup Contact. 2014.

136. Batson CD, Fultz J, Schoenrade PA, Paduano A. Critical self-reflection and self-perceived altruism: when self-reward fails. J Pers Soc Psychol. 1987;53: 594–602.

137. Batson CD, Fultz J, Schoenrade PA. Distress and empathy: two qualitatively distinct vicarious emotions with different motivational consequences. J Pers. 1987;55: 19–39.

138. Hogan R. Development Of An Empathy Scale. J Consult Clin Psychol. 1969;33: 307–316.

139. Gough HG. A leadership index on the California Psychological Inventory. J Couns Psychol. 1969;16: 283–289.

140. Mehrabian A, Epstein N. A measure of emotional empathy. J Pers. 1972;40: 525–543.

141. Long ECJ. Measuring Dyadic Perspective-Taking: Two Scales for Assessing Perspective-Taking in Marriage and Similar Dyads. Educ Psychol Meas. 1990;50: 91–103.

142. Axtell CM, Parker SK, Holman D, Totterdell P. Enhancing customer service: Perspective taking in a call centre. Eur J Work Org Psychol. 2007;16: 141–168.

143. Cabrera L, Cabrera D. The Systems Thinking & Metacognition Inventory (STMI). Cabrera Research Lab; 2015.

144. Taylor CS. Validity and Validation. Oxford University Press; 2013.

145. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol. 1999;77: 1121–1134.