Formative Evaluation of Students at Quest to Learn

IJLM

Valarie Shute
Florida State University
vshute@fsu.edu

Matthew Ventura
Florida State University
mventura@fsu.edu

Robert Torres
The Bill & Melinda Gates Foundation
Robert.Torres@gatesfoundation.org

Keywords
assessment
evidence-centered design
game-based pedagogy

Abstract

Quest to Learn (Q2L) is a public school that opened in 2009 to respond to the needs of kids who are growing up in a highly networked, globally competitive, and complex world. The school was designed to include a game-based pedagogy that provides a dynamic, interdisciplinary curriculum, as well as immersive, challenge-based learning experiences. The present article reports the findings from a 20-month evaluation of Q2L. This research was launched when the school opened, with the main goal to identify and assess a set of key student competencies that are relevant and important to Q2L. This student-level evaluation focused on students' development of three core skills—systems thinking, time management, and teamwork—over the 20-month time period. We found that students significantly improved their systems thinking skills over the duration of the study, and improved (albeit, not significantly) on their time management and teamwork skills as well.

Visit IJLM.net

doi:10.1162/IJLM_a_00087

Volume 4, Number 1

Formative Evaluation of Student Learning at Quest to Learn

> Much education today is monumentally ineffective. All too often we are giving young people cut flowers when we should be teaching them to grow their own plants.
>
> —J. W. Gardner (1995, p. 21)

Children growing up in the 21st century need to be able to think differently from kids in past generations. When confronted with complex problems, they need to be able to think creatively, systemically, critically, as well as work collaboratively with others (see, e.g., Partnership for 21st Century Skills 2006; Resnick 2007). In contrast to these needs, current educational practices typically adhere to outdated theories of learning and pedagogy, evidenced by a so-called content fetish (Gee 2005). That is, schools focus on increasing students' proficiency in traditional subjects such as math and reading, via didactic approaches, which leaves many students disengaged. Schools that depart from the entrenched pedagogy are the exception to the rule.

Quest to Learn (Q2L) is one of these exceptions. Q2L was designed to respond to the needs of kids who are growing up in a digital, information-rich, globally competitive, and complex world. Designed to serve students in grades 6–12, the school (as of the 2011–2012 school year) houses about 240 students in grades 6–8, and will add one grade each year until full capacity is reached in 2015 with approximately 560 students. Q2L opened in 2009 as a regular (noncharter) New York City Department of Education public school in Manhattan's Community School District 2. Students are selected through an open lottery system. Admission is not based on prior academic achievement. A sister school to Q2L, called ChicagoQuest, which uses the same game-based curricular model, opened in Chicago in 2011. Both schools are intended to enable all students, regardless of their academic or personal challenges, to contribute to the design and innovation necessary to meet the needs and demands of a global knowledge society. The model provides a learning environment that includes a dynamic, interdisciplinary curriculum, as well as immersive, challenge-based learning experiences (Salen et al. 2011). Toward that end, the model's culture and pedagogy aim to foster deep curiosity for lifelong learning as well as a commitment to social responsibility and respect for others and self. For more, see http://q2l.org/.

Background: Q2L Curriculum and Assessment

The Q2L curriculum is inspired by the notion that learning is not simply an individualized affair but is a highly social, context-dependent, and collaborative achievement (see, e.g., Vygotsky 1978; Lave and Wenger 1991; Bransford, Brown, and Cocking 2000). Q2L uses a systems-thinking framework as a core curricular and pedagogical strategy. Systems thinking is defined as a holistic perspective that sees the world as increasingly, progressively interconnected, from elemental components to complex systems of activity (see, e.g., Forrester 1994; Salisbury 1996; Ossimitz 2000; Assaraf and Orion 2005; Barak and Williams 2007). This broad definition includes social, natural, and technological systems that can be studied and understood as having certain cross-cutting commonalities, such as rules, goals, and particular behaviors.

The school also employs a game-based pedagogy, designed by teachers, professional game designers, curriculum directors, and other content experts. Collectively, they create a series of 10-week "missions" (i.e., units of study) that constitute an immersive world in which students take on various identities—from cartographers to architects to zoologists—to solve design and systems-based problems.

Beyond simply employing games to teach, Q2L uses the internal architecture of games to create game-like learning environments, whether analog or digital. Games instantiate constrained systems of activity (or worlds) in which players engage mediational tools, other players, and rule sets to achieve clearly defined winning (or goal) conditions. This requires players to enact specific game behaviors. These behaviors in the gaming world are called core mechanics, which players must perform to successfully complete a game. In the classroom, similar behaviors are required to successfully acquire new knowledge and skills.

To suit the needs of the 21st century, learning environments should reflect knowledge domains (or discourse communities) that reflect the epistemology of current real world professional and/or industry domains. That is, learners should be offered opportunities to apply the actual kinds of knowledge, skills, values, and behaviors (i.e., core mechanics) needed to participate in the domain of, say, U.S. history, journalism, or cellular biology. Learners should be

Table 1 Three learning dimensions of Q2L

Dimension 1: *Civic & Social-Emotional Learning*	Dimension 2: *Design*	Dimension 3: *Content*
Learning from peers and others	Systems thinking	Code worlds
Planning, organizing, adapting, and managing goals and priorities	Iteration	Wellness
Persisting to overcome complex challenges	Digital media tool use	The way things work

offered genuine and ample opportunities to produce and iterate on content endemic to real knowledge domains. Thus, rather than working with the cut flowers provided to them by traditional pedagogy, students should be learning to grow their own flowering plants. Q2L strives to meet these goals. Moreover, Q2L students are provided with (and often establish on their own) communities of practice in which they can collaborate and informally (or formally) share their work with a community of peers (as well as with teachers and game and curriculum designers, as warranted). Similarly, the game-like curriculum at Q2L is intended to engage students in collaborations with others to complete various tasks. Many quests (i.e., smaller units of a mission) are designed to allow students to work with one another. These quests require the integration of different areas of expertise and also allow students to provide one another with ongoing feedback.

The school also employs a distinctive approach to assessment. The basic principle at Q2L is that assessment should be embedded within learning and should take place in situ—located in the discourse and actions of students and teachers (Salen et al. 2011). Therefore, students' knowledge and performance is assessed at Q2L by holistic and qualitative techniques such as interviews and observations. Q2L's original design documents outline three learning dimensions—civic and social-emotional learning, design, and content—that guide the curriculum and assessment.

Table 1 (adapted from Salen et al. 2011) depicts some of the specific and valued competencies that undergird and frame Q2L. Assessment in Q2L does not focus only on how much content a student knows but on how she *uses* what she knows while displaying other important 21st-century competencies (e.g., systems thinking, teamwork, and time management skills).

Research Approach

Q2L redefines what a school looks and feels like and what kids need to learn to succeed—at school and in life. Ultimately, Q2L may be able to provide a new school model for 21st-century learning. Before the model is applied more broadly, however, we need to answer some important questions. For instance, how do we define what constitutes "success" at Q2L? How do we know if (and to what degree) Q2L is accomplishing its goals? Our 20-month study (September 2009–May 2011) was implemented as a preliminary attempt to answer these questions. The answer to the first question, concerning the identification of important school-level variables, is described elsewhere (Shute and Torres 2012). Our second question examines growth in focal student-level variables and is the focus of the present article.

We set up our study to coincide with the school opening in September 2009. In doing so, our main research goal was to identify and assess a set of key student competencies that are relevant and important to Q2L.

Student-Level Evaluation

The purpose of this article is to report the results of our assessment of specific 21st-century competencies that were identified as important to the school. We periodically assessed the following three competencies during the 20-month period: (1) systems thinking skill, (2) teamwork (i.e., learning from peers and others), and (3) time management skill (i.e., planning, organizing, and managing goals and priorities).

We selected these three competencies because (a) they are identified in the original Q2L design documents as core competencies that the school intends to support (see Table 1), (b) they were consistently identified by Q2L stakeholders/interviewees and are included in the school objectives model (see Shute and Torres 2012), and (c) validated assessment instruments for these three competencies already exist. Each of these competencies has its own set of facets. Figure 1 shows the three competencies and their facets.

Our assessment for the systems thinking skill (Figure 1a) is based on the protocol employed by

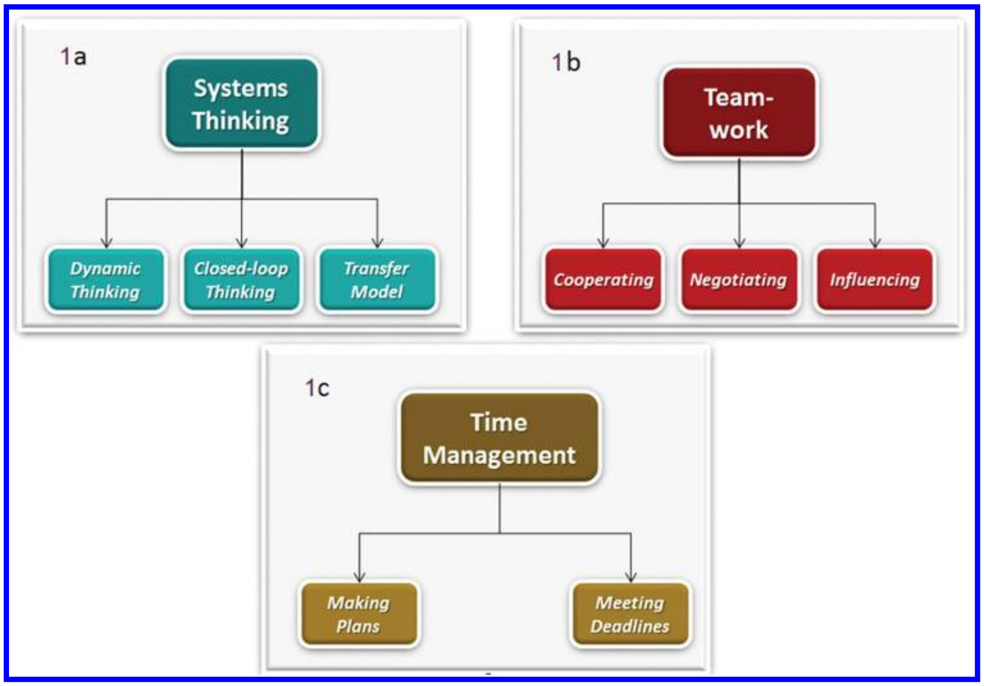


Fig. 1 Three competency models for student assessment at Q2L

Sweeney and Sterman (2007). For the teamwork model (Figure 1b), we synthesized research described by Rysavy and Sales (1991); Totten et al. (1991); Van den Bossche, Segers, and Kirschner (2006); Tindale, Stawiski, and Jacobs (2008); and Zhuang et al. (2008). The three-factor solution shown in Figure 1b has been consistently reported for this variable relative to middle- and high-school students (see Zhuang et al. 2008). Finally, our time management model (Figure 1c) was based on the findings of Macan et al. (1990); Roberts, Schulze, and Minsky (2006); Liu et al. (2009); MacCann, Duckworth, and Roberts (2009). This two-factor solution has been reported relative to middle-school students.

The first of our research questions related to student assessment was: *Does the incoming cohort of sixth-grade students at Q2L demonstrate improved performance over 20 months in the areas of systems thinking, teamwork, and time management skills?* These three competencies are supported (albeit implicitly) during the course of daily school activities (e.g., completing projects in small groups, designing games).

In addition, we examined this cohort's academic achievement (i.e., mathematics and reading skills) in relation to New York State standardized tests. Thus, the second research question is: *Do the three focal competencies predict academic achievement relative to the state mandated math and reading test scores?* We tested whether Q2L was successfully supporting students' development of important new competencies as a function of its unique environment, while not sacrificing traditional (i.e., math and reading) academic achievements.

For our first research question, we hypothesized that students would, on average, demonstrate improved performance in relation to the three focal competencies from the initial (Time 1, September 2009) to the final (Time 4, May 2011) assessment. For research question 2, we hypothesized that there would be significant correlations among our three competencies and the annual standardized state test scores, despite the fact that Q2L offers no explicit math or reading courses. Courses at Q2L are interdisciplinary in nature—math and reading skills, along with systems thinking and design thinking skills, are exercised in courses with titles such as Code Worlds and The Way Things Work.

We used existing instruments to assess our three focal competencies during this research project. The systems thinking competency was assessed using a modified version of the protocol described in Torres's (2009) dissertation on the topic, which in turn was based on the protocol described in Sweeney and Sterman (2007). This assessment contains 12 constructed-response items, with different examples of systems used across administrations (e.g., hunger/eating versus predator/prey relations). Appendix 1 includes the full set of questions used in the systems thinking assessment.

Teamwork and time management skills were measured via instruments that have been designed, developed, and validated by the Educational Testing Service for use by middle-school students (see, e.g., Zhuang et al. 2008; MacCann, Duckworth, and Roberts 2009; Wang et al. 2009). The first teamwork assessment contained 57 Likert-scale items and 12 scenario-based items, but subsequent administrations of the assessment used fewer items (i.e., the 22 Likert-scale items that had the strongest weights on the three focal factors—cooperating, negotiating, and influencing others—along with 12 scenario-based items). The first time management assessment contained 36 Likert-scale items, and subsequent administrations employed 24 items (again, those weighting most strongly on the two focal factors: making plans and meeting deadlines). The items making up the assessments for both the teamwork and time management skills are presented in Appendices 2 and 3.

Example items (with text boxes for entering constructed responses) from the systems thinking protocol include: "How are hunger and eating related or connected to each other? If you just finished eating, what happens to your level of hunger over time? Can you think of another situation that feels or seems like the same as this?" Rubrics were adapted from Torres's (2009) and Sweeney and Sterman's (2007) research to score the responses. For instance, the systems thinking rubric has five levels (from 0 to 4). Each level contains a description as well as multiple examples of student responses for each of the three main variables: dynamic thinking, closed-loop thinking, and transfer of models (i.e., Level 0: Incorrect or nonapplicable response; Level 1: Describes static interconnections; Level 2: Describes aspects of system structures and behaviors; Level 3: Demonstrates understanding of principles guiding system behaviors [though descriptions may be limited]; and Level 4: Full utilization of systems intelligence, such as a description of a system at multiple levels). See Table 2 for the five scoring levels across the three main facets of ST.

Example items from the teamwork survey (with five-point Likert-scale responses, from "never" to "always") include "I don't have an opinion until all of the facts are known"; "I know how to make other students see things my way"; "I give in when arguing"; and "I find it difficult to keep team members on task." Example items from the time management survey (also on a five-point Likert scale, from "never" to "always") include "I complete my homework on time"; "I put off tasks until the last minute"; "I keep my desk neat"; and "I like routine." Prior to data analysis, some items were reverse coded so that all items would be in the same direction (i.e., higher values equate to more often or more positive).

The student assessments occurred at approximately six-month intervals,[1] including the initial and final weeks of the 20-month period, yielding four data collection times (September 2009, March 2010, November 2010, and May 2011). This enabled us to capture current competency levels and monitor the development of competencies over time. The systems thinking assessment took approximately 45 minutes, while the teamwork and time management assessments took around 25 minutes each to complete. All assessments were administered during nonacademic periods and were staggered across two days (i.e., systems thinking on one day and teamwork and time management on the next) to attenuate fatigue. The first administration of each of the three assessment instruments was delivered in paper and pencil format. The three remaining administrations were delivered online, where students clicked on a link that took them directly to the assessment. Earlier analyses showed that delivery mode had no effect on assessment outcomes (Shute and Torres 2012).

Traditional achievement testing occurred in June 2010 as part of normal New York State requirements for sixth-grade students.

Method

Sample

The primary sample for our study consisted of 70 Q2L students (42 male, 28 female), beginning in their sixth-grade year (2009–2010) and concluding at the end of their seventh-grade year (2010–2011). They were assessed four times. We also tested students who entered Q2L in the fall of 2010 as new sixth graders ($n = 76$). This secondary cohort was assessed only twice.

At the time of writing, about 40% of Q2L students received free or reduced lunch, and 60% paid for lunch in full. Approximately 34% of the students were Caucasian, 32% were African-American, 29% were Hispanic, and 5% were Asian.

Table 2 Systems thinking scoring rubric (adapted from Sweeney and Sterman 2007)

	Dynamic Thinking Level A: Identification of System Elements and Interrelations Level B: Specify Variables and Problems	Closed-Loop Thinking Level A: Understanding of Basic System Concepts Level B: Elaborate Causal Reasoning	Transfer the Model Identification of Similarities across Systems
Level 0: *Incorrect or nonapplicable response*	No response, response of "I don't know," or nonapplicable response.	No response, response of "I don't know," or nonapplicable response.	No response, response of "I don't know," or nonapplicable response.
Level 1: *Describes static interconnections*	Includes identification of *some* discrete elements within a system. Interconnections and interrelationships are described in *linear and/or static* (vs. dynamic) terms.	Includes a description of interconnections and interrelationships in linear and/or static (vs. in closed loop) terms.	Involves an attempt to indicate a similar structure.
Level 2: *Describes aspects of system structures and behaviors*	Shows *some* identification of discreet elements within a system, understanding of the attributes and characteristics of elements within a system, including a dynamic or an interconnection.	Shows *some* understanding of the behaviors and characteristics of reinforcing *and/or* balancing feedback loops, including looped interconnections, time delays, and feedback types.	Shows *some* understanding of the similarities and differences between natural, technological, or social systems.
Level 3: *Demonstrates understanding of principles guiding system behaviors (though descriptions may be limited)*	Demonstrates *sound* understanding of discreet elements within a system, understanding of the attributes and characteristics of elements within a system, including a dynamic or an interconnection.	Demonstrates *sound* understanding of the behaviors and characteristics of reinforcing *and/or* balancing feedback loops, including looped interconnections, time delays, and feedback types.	Demonstrates *sound* understanding of the similarities and differences between natural, technological, or social systems.
Level 4: *Full utilization of systems intelligence, such as a* description *of a system at multiple levels*	Includes a *full manifestation* of understanding of discreet elements within a system and understanding of the behaviors and characteristics of elements within a system, including *multiple dynamic* interconnections *at multiples levels*.	Includes a *full manifestation* of systemic reasoning, including understanding of the behaviors and characteristics of reinforcing and/or balancing feedback loops and a description of looped interconnections, time delays, and feedback types.	Includes a clear understanding of the similarities and differences between natural, technological, or social systems and includes *clear and novel* observations and descriptions of similar structures.

Reliability

Before analyzing student performance on our set of assessments, we had to establish that the instruments were sound (i.e., reliable and valid). Because the systems thinking assessment was the only one involving constructed responses, we additionally needed to determine the inter-rater reliability of the scores.

Inter-rater Reliability of Systems Thinking Scores

To score the responses on the systems thinking assessment, we used two separate teams of scorers, with two persons per team. Scoring in the first administration (Time 1) was accomplished by the two teams independently scoring all students' constructed responses using our five-point (levels 0–4) rubric. Both teams (a) recorded their scores in an Excel spreadsheet, then (b) exchanged spreadsheets with the other team, and (c) highlighted scores that differed by more than two points. The highlighted scores were then discussed and rescored. In most cases the rescored responses converged on the same number or differed by only one point. In the first administration of the systems thinking assessment, this exchange-discuss-revise

Table 3 Reliabilities (and sample sizes) for assessments across four time periods

Competency	Time 1	Time 2	Time 3	Time 4
Systems thinking	0.82 ($n = 70$)	0.85 ($n = 66$)	0.80 ($n = 41$)	0.84 ($n = 43$)
Teamwork	0.84 ($n = 63$)	0.82 ($n = 59$)	0.88 ($n = 49$)	0.86 ($n = 53$)
Time management	0.81 ($n = 69$)	0.83 ($n = 62$)	0.84 ($n = 59$)	0.84 ($n = 54$)

Table 4 Focal competency means, SDs, and statistics for the four administrations

	Time 1	Time 2	Time 3	Time 4	*n*	*F*
Systems thinking	0.90 (0.42)	1.20 (0.53)	0.95 (0.48)	1.13 (0.62)	26	2.07
Teamwork	3.49 (0.45)	3.40 (0.41)	3.60 (0.45)	3.51 (0.29)	33	1.03
Time management	3.65 (0.52)	3.57 (0.49)	3.55 (0.50)	3.65 (0.45)	35	0.01

process required two full iterations until no scores were more than one point different. For the second, third, and fourth administrations, which used the same teams, just one iteration was required, with only a few discrepant scores requiring discussion. The inter-rater reliabilities of the scores on the systems thinking assessment were: Time 1, Spearman's rho = 0.87; Time 2, Spearman's rho = 0.96. Subsequent administrations, using the same two-team process, showed similarly high inter-rater reliabilities (i.e., in the mid-90s).

Internal Reliability

The alpha reliabilities of the three assessments were calculated across all four administrations and were judged to be appropriate. See Table 3 for Cronbach alpha reliabilities for each of the three assessments across the four time periods.

Results

Table 4 displays the overall means (and standard deviations) of the three competencies over the four time points.

We computed a repeated measure ANOVA using Time 1, Time 2, Time 3, and Time 4 as the repeated measures. As table 4 shows, no significant differences were found among the means across all four time points. However, the means in table 4 represent students who participated in *all four* administrations. That is, if a student missed even one administration (e.g., completed assessments at Times 1, 3, and 4, but was absent at Time 2), her data (from all four times) were removed from the analysis because of our repeated-measures design. In addition, data collection with middle-school students can be difficult. Assessments were given in the classroom, where students were sitting adjacent to one another. This created distractions for students who were trying to take the assessment seriously. Because of these issues, we felt that a better analysis was to take the first and fourth administrations and run a separate analysis to see whether this led to any significant changes to mean scores.

Table 5 Competency means, SDs, and statistics for the first and fourth administration

	Time 1	Time 4	*n*	*t*
Systems thinking	0.95 (0.43)	1.21 (0.57)	35	3.72*
Teamwork	3.46 (0.46)	3.53 (0.33)	42	0.97
Time management	3.55 (0.55)	3.59 (0.49)	46	0.72

*$p < .01$

This simplified analysis also serves to increase the sample size (compared to the data in Table 4) because it includes students who participated in at least the first and fourth administrations rather than all four administrations. Thus, for the systems thinking assessment, the sample size of students who participated in all four administrations (shown in Table 4) was $n = 26$, whereas the sample size of those who completed the Time 1 and Time 4 assessments (shown in Table 5) was $n = 35$. Table 5 displays the means of the three competencies for students who took the assessments at Time 1 and Time 4.

The results show a significant improvement from Time 1 to Time 4 for the systems thinking assessment. No significant improvements were observed for teamwork and time management.

Systems Thinking (ST)

As shown in Table 5, students who started at Q2L as sixth graders when the school opened in September 2009 significantly improved on their overall ST skills from Time 1 to Time 4, when they were finishing seventh grade ($t\ (34) = 3.72$; $p < .001$). When the three

Table 6 Correlations among ST measures across the four time points

	ST2	ST3	ST4
ST1	0.51**	0.55**	0.69**
ST2		0.37*	0.57**
ST3			0.74**

*$p < .05$; **$p < .01$

Table 7 Correlations among TW measures across the four time points

	TW2	TW3	TW4
TW1	0.53**	0.47*	0.43*
TW2		0.70**	0.55**
TW3			0.57**

*$p < .05$; **$p < .01$

Table 8 Correlations among TM measures across four time points

	TM2	TM3	TM4
TM1	0.68**	0.34*	0.59**
TM2		0.66**	0.63**
TM3			0.74**

*$p < .05$; **$p < .01$

Table 9 Correlations among competencies at Time 4

	TM	TW
ST	0.30	0.15
TM		0.54*

*$p < .01$

ST facets were analyzed separately (see Figure 1a), we found significant gains for two of them—dynamic thinking and transfer the model (as well as a non-significant gain for closed-loop thinking). The second cohort of students (i.e., those who started sixth grade in September 2010) also showed significant gains on their ST skills from Time 1 ($M = 0.58$; $SD = 0.24$) to Time 2 ($M = 0.88$; $SD = 0.36$); ($t\,(21) = 4.20$; $p < .001$).

The following responses by a Q2L student illustrate growth in ST skills from Time 1 to Time 4. In response to (a) “What is an example of a system?” and (b) “Identify at least two parts of your system and explain how they relate to each other,” the student at Time 1 responded, (a) “An example is a puzzle”; (b) “I don’t know.” At Time 4 her responses were more meaningful: (a) “The transportation or subway system is an example of a system”; (b) “Two parts of the transportation system are the people and the metrocards. They relate to each other because the system wouldn’t work if people didn’t buy metrocards to use for the trains. If there were no metrocards, people wouldn’t be able to go on the trains and buses, and the government would not be able to make a profit.”

Table 6 displays the relationship among ST scores across the four time points and shows that the test-retest reliability of the ST assessment is consistent over the four time points.

Teamwork (TW)

Although students showed higher Time 4 than Time 1 scores on their overall TW scores, the difference was not significant. However one of the facets of TW, influencing others, did show a statistically significant increase over time ($t\,(41) = 5.13$; $p < .001$). The sixth graders (i.e., the second cohort) did not show significant gains on TW—either overall or by factors. Table 7 displays the relationship among TW scores across the four time points and shows that the test-retest reliability of the TW assessment is consistent over the four time points.

Time Management (TM)

Similar to the findings of TW over time, students showed a small improvement on their overall TM scores from Time 1 to Time 4, but the improvement was not statistically significant. However, one of the TM facets, meeting deadlines, did show a significant gain ($t\,(45) = 2.18$; $p < .05$). The 6th graders (i.e., the second cohort) did not show significant gains on TM—either overall or by factors. Table 8 displays the relationship among TM across the four time points and shows that the test-retest reliability of the TM assessment is consistent over time.

Relationships among the Three Competencies

Table 9 displays the relationships among the three competencies at Time 4. No significant relationship was found between ST and either TW or TM. We did, however, find a moderate relationship between TM and TW. This shows that the three assessments are not measuring the same competencies. That is, if the correlation between TW and TM had been 0.70 or higher (it was not), we would suspect the assessments were measuring the same construct. Also, because TW

Table 10 Hierarchical regression analysis with competencies and math

Model	R-Square	F Change
1—ST	0.219	9.80*
2—ST, TW	0.235	0.74
3—ST, TW, TM	0.236	0.03

*$p < .01$

Table 11 Hierarchical regression analysis with competencies and reading

Model	R-Square	F Change
1—ST	0.163	7.21*
2—ST, TW	0.194	1.36
3—ST, TW, TM	0.238	2.02

*$p < .05$

and TM were assessed via the same method (i.e., self-report, with five-point Likert scale), we expected some degree of overlap because they share method variance.

Predictive Validity

We found that the scores from the last administration of the three competency assessments correlated with state test scores. ST significantly correlated with math ($r = 0.44$, $p < .01$) and reading ($r = 0.41$, $p < .01$) test scores; TW significantly correlated with math ($r = 0.29$, $p < .05$) but not reading; and TM marginally correlated with reading ($r = 0.27$, $p = .05$) but not math. Scores came from the final (Time 4) assessments. This suggests that the competencies are positively associated with state test scores.

To examine the incremental validity of the three competencies on state standardized math and reading scores, we conducted two hierarchical regression analyses. In the first, we predicted math scores from ST, TW, and TM (see Table 10). We chose to enter ST first in both analyses because it was a more cognitive measure than either TW or TM.

Table 11 shows the results from our hierarchical regression analysis predicting reading test score.

Results from both regression analyses indicate that ST significantly predicts both math (22% unique variance) and reading (16% unique variance) standardized test scores, while TW and TM do not appear to predict any unique variance of the state test scores. This suggests that (1) ST is an important skill to

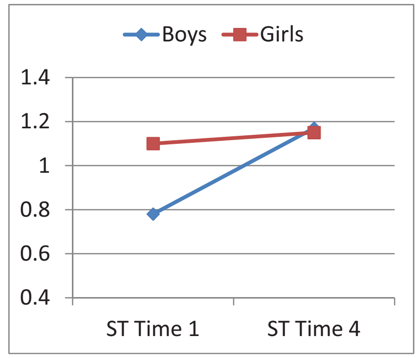


Fig. 2 Interaction between ST (Time 1 versus Time 4) and gender

continue to support because it positively predicts state test scores, and (2) while TW and TM might not predict state test scores, we may still want to explore alternative methods of assessment for TW and TM.

Gender and Socioeconomic Status

Although the sample size is small, we conducted an exploratory examination of the effects of gender and socioeconomic status (SES) in relation to ST growth. First, we separately examined boys' ($n = 21$) and girls' ($n = 16$) scores for ST at Time 1 versus Time 4. We computed a repeated measures ANOVA (ST Time 1 and Time 4) with gender as the between-subjects variable. The main effect was significant (F (1, 35) = 10.42; $p = .003$) and the interaction with gender was also significant (F (1, 35) = 5.99; $p = .02$). A graph of the interaction is shown in Figure 2. Further analysis showed that the boys significantly improved over time in ST, while the girls did not (t (20) = 4.10; $p = .001$).

We similarly evaluated gender and our other two competencies (i.e., TW and TM) but found no interactions.

Our SES variable was coded relative to whether students at Q2L received a free or reduced lunch. We computed a repeated measures ANOVA for our three competencies (at Time 1 versus Time 4) with SES as the between-subjects variable. For each of the three competencies, no interactions involving SES were found. This may be viewed as a sign that students, regardless of SES, appear to be improving equally on ST.

Discussion

This research attempted to answer preliminary questions regarding the effectiveness of Q2L, particularly in relation to supporting important 21st-century competencies. We hypothesized that students would demonstrate improved performance in relation to the three focal competencies from initial (Time 1, September 2009) to final (Time 4, May 2011) assessment. We found partial support for this hypothesis: systems thinking did improve significantly when comparing the data from Time 1 and Time 4. While we did not find growth in teamwork and time management from Time 1 to Time 4, this might be attributable to the method (i.e., self-report) by which we chose to assess these competencies. Self-report measures are easy to administer, but they are not typically used to measure growth over time. Additionally, self-report measures are susceptible to social desirability effects that can affect the validity of the assessment. Using methods that are more performance-based (e.g., classroom observation) to assess time management and teamwork in Q2L might yield more favorable results.

Although the results of this formative evaluation are interesting—particularly regarding the improvement for systems thinking skill from Time 1 to Time 4—more research is needed to understand fully how (and why) Q2L impacts students. Because of the relatively small sample size, the growth of competencies could not be unequivocally measured. In the future, larger samples of students will allow researchers to see how certain student characteristics and demographics (e.g., gender, socioeconomic status) as well as context play a role in how Q2L impacts students. For example, how do girls versus boys grow in important competencies in Q2L over time? Does Q2L help to reduce the achievement gap among different SES brackets or ethnicities? What are the best settings at Q2L to foster better thinking (and behaving) skills?

Given the current design, we do not conclusively know whether students' gains are simply a function of normal cognitive development or are, in fact, a function of being part of Q2L. Furthermore, the current research did not include a control group in the analysis of competency growth. However, because students at Q2L did demonstrate significant improvements in systems thinking skill across the 20-month period and because this skill is difficult to master (even for adults), we suspect that the systems thinking growth was primarily a function of being part of Q2L.

While the assessments that we used in this study were adequate for the exploratory nature of this research, they are not optimal for large-scale, meaningful evaluations within Q2L and across multiple schools—especially for the teamwork and time management competencies. Future research should focus on using performance-based assessments across a range of competencies that are of value to Q2L.

Finally, this evaluation research should be viewed as a preliminary probe of the opportunities and challenges encountered in the process of translating a new, innovative model of schooling into practice. As such, we have analyzed only the tip of the proverbial iceberg that is the Q2L model, with its many dynamic and interrelated parts (e.g., interdisciplinary curriculum, immersive environment, challenge-based learning experiences). As with any bold new idea, the success of its implementation will not be known for years, and certainly not after just 20 months.

Appendix 1—Systems Thinking Assessment Questions

Directions

Below you will find 12 questions that you should read and respond to. Please answer *all* questions. Don't spend too much time deciding on what your answer should be. There are no right or wrong answers.

SYSTEMS

1. *What does the word "system" mean to you?*
2. *What is an example of a system?*
3. *Identify at least two parts of the system and explain how they relate to each other.*

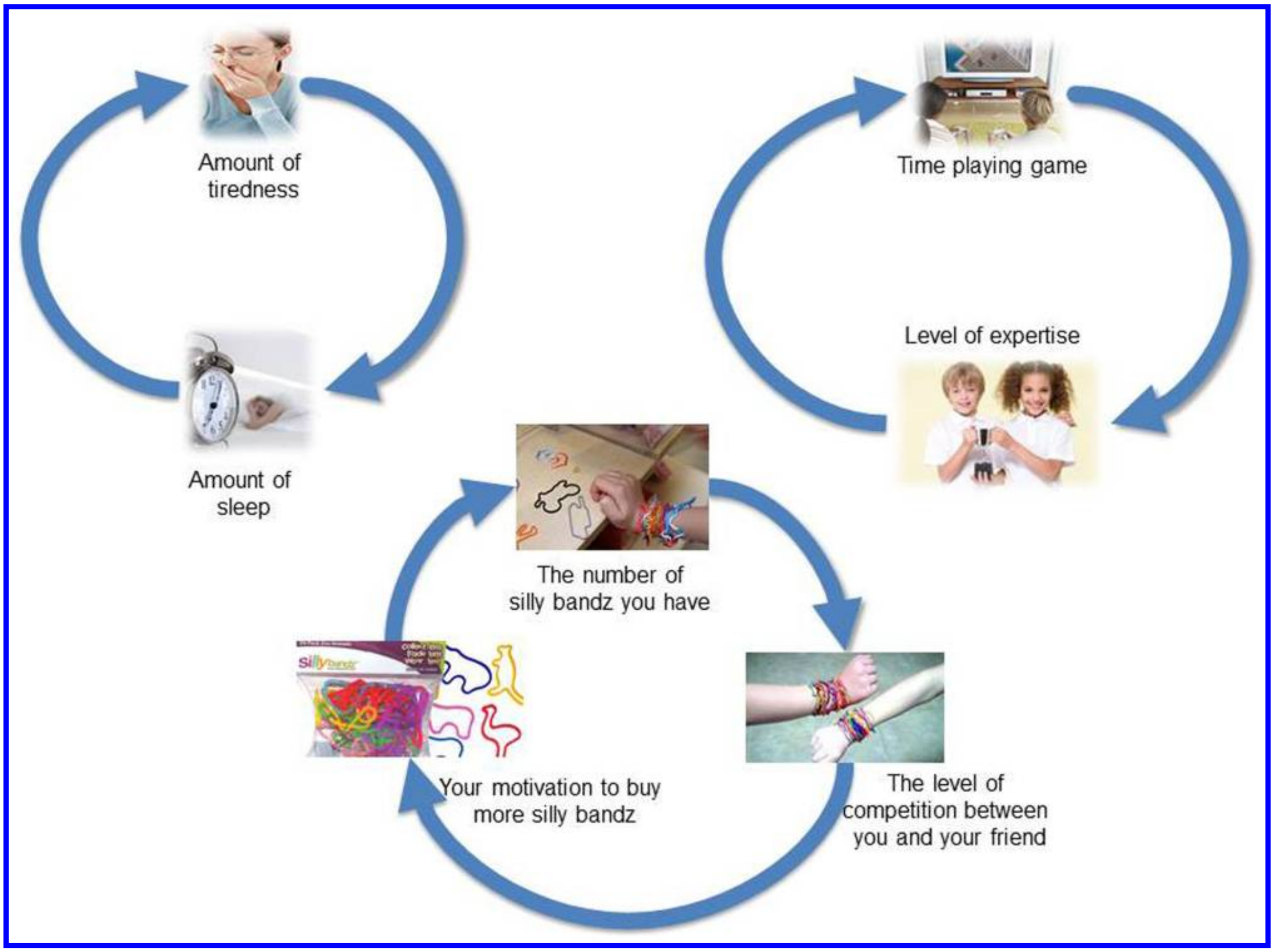
TIREDNESS AND SLEEP

Think about the relationship between tiredness and the number of hours slept.

4. *How are tiredness and amount of sleep related or connected to each other?*
5. *If you sleep less, what happens to your level of tiredness over time?*
6. *Can you think of a different situation that feels the same as this?*

SILLY BANDZ

7. *You and your best friend Margot have been collecting silly bandz for a year. You often argue about who has more. You definitely want to have more silly bandz than Margot because you don't want her to brag about it. Think about the relationship(s) among: (a) the*



number of silly bandz you have, (b) the level of competition between you and Margot, and (c) your motivation to buy more silly bandz. How might these three be interrelated? What do you think happens to your motivation to buy more silly bandz and the number of silly bandz you have over time?

8. *If you have more silly bandz, what happens to the level of competition between you and Margot and your motivation to buy more silly bandz?*
9. *Can you think of a different situation that feels the same as this?*

PLAYING GAMES

10. *How does the amount of time spent playing your favorite game affect your expertise in playing the game? What do you think happens to these two things over time?*
11. *Can you think of a different situation that feels the same as this?*

CYCLES

12. *Take a look at the image above with the three cycles (tiredness, games, and silly bandz). Which two do you think are similar and why?*

Appendix 2—Teamwork Survey

Section 1: Thinking, Feeling, and Acting

Directions

Below you'll find 22 statements about how you behave when you are in team situations or during group projects. You should read and decide how well each one of them describes the way you think or feel in these situations. For each item, please choose one of the five options ranging from "never" to "always." Please reply to all statements. Don't spend too long deciding what your answer should be. Answer all statements even if you're not entirely sure of your answer. There are no right or wrong answers. SELECT

THE CIRCLE FOR EACH ITEM THAT MATCHES YOUR BEST RESPONSE.

1. *I don't have an opinion until all of the facts are known.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

2. *I know how to make other students see things my way.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

3. *I am flexible when doing group projects.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

4. *I dislike being in a position of responsibility for other people.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

5. *I like to be in charge of group projects.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

6. *I enjoy helping team members.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

7. *I prefer working alone over working with others.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

8. *I can make deals in any situation.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

9. *Feedback is important to me.*

   Never Rarely Sometimes Often Always
   ❍ ❍ ❍ ❍ ❍

10. *I like to solve problems using different tactics.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

11. *During group assignments, I make demands on other students.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

12. *I suggest different solutions to problems.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

13. *I find it difficult to approach classmates.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

14. *I enjoy bringing team members together.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

15. *I believe that I'm a good leader.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

16. *I can convince my peers about anything.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

17. *I can fight for a cause that I believe in.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

18. *I know when to step in when an argument starts getting out of control.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

19. *I'm influenced by other students' opinions.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

20. *I learn from other students.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

21. *I think that trading ideas among students leads to the best solutions.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

22. *I am inspired by others' ideas and thoughts.*

    Never Rarely Sometimes Often Always
    ❍ ❍ ❍ ❍ ❍

Section 2: Scenarios (Optional)

Directions

In this section, you'll find a number of paragraphs describing various situations (scenarios). You need to evaluate how effective you think several reactions to the situation are. Please read each scenario carefully before choosing your response. Don't spend too long deciding on each answer. Please answer all of the statements even if you're not entirely sure of your answer. There are no right or wrong answers.

SCENARIO 1. You're part of a study group that has been assigned a large presentation for class. As you are all dividing up the work, it becomes clear that both you and another team member are interested in researching the same aspect of the topic. Your friend already knows a lot about the area, but you have been extremely excited about working on this part of the assignment ever since your teacher mentioned it.

1. *Flip a coin to determine who gets to work on this part of the assignment.*

   Very Ineffective Ineffective Neutral Effective Very Effective
   ❍ ❍ ❍ ❍ ❍

2. *Insist that, for the good of the group, you should work on this part of the assignment because your interest in the area means you will do a really good job.*

   Very Ineffective Ineffective Neutral Effective Very Effective
   ❍ ❍ ❍ ❍ ❍

3. *Forget all about your interest and allow your friend to work on this aspect of the assignment.*

   Very Ineffective Ineffective Neutral Effective Very Effective
   ❍ ❍ ❍ ❍ ❍

4. *Suggest to the other group member that you both share the research for this aspect of the assignment. You can then also share the research on another less desirable part.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

SCENARIO 2. Imagine that your teacher has assigned each person in the class to a small group to work on a problem. After breaking into small groups to discuss the problem, everyone gets back together in class to go over the ideas generated in each small group. When it's your group's turn, the person who volunteered to serve as spokesperson begins to present his or her own ideas, some of which had not been discussed in the small group, rather than the ideas generated in the group discussion.

1. *Do nothing in order to preserve the impression that your group worked well together.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

2. *Challenge the spokesperson, because it's important that only the ideas agreed upon by your group be presented.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

3. *Keep silent during class, but afterwards speak with the instructor privately about the spokesperson.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

4. *Politely add additional comments, making sure that the ideas the group discussed are presented.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

SCENARIO 3. You've decided to set up a new website about your school—from the perspective of students. You've been given permission and a deadline: three weeks to complete it. Nine other students have signed up to help you. At your first meeting, you want to assign jobs for everyone, such as creating the graphics (2D and 3D), animations, code, content, online surveys/polls, and so on. But how will you assign each person to the tasks?

1. *Ask each student which job they would like and then resolve any conflicts based on who asked first.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

2. *Ask each team member which job they would like to take on, and then resolve any conflicts by drawing names out of a hat.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

3. *Make a list of the team members' strengths and preferences and assign jobs that match their strengths most closely.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

4. *Assign the best jobs to your friends first, and then give the remaining jobs to the rest of the team members randomly.*

Very Ineffective	Ineffective	Neutral	Effective	Very Effective
❍	❍	❍	❍	❍

Appendix 3—Time Management Survey

Directions

Below you will find a number of statements that you should read and decide how well each one of them describes the way you think or feel. *Please reply to all statements*. Give your first impression of whether each statement describes the way you think and feel about it. Don't spend too long on deciding what your answer should be. Answer all statements even if you're not entirely sure of your answer. There are no right or wrong answers. SELECT THE CIRCLE FOR EACH ITEM THAT MATCHES YOUR BEST RESPONSE.

	Item	Never	Rarely	Sometimes	Often	Always
1.	*I complete my homework on time.*	❍	❍	❍	❍	❍
2.	*I never put off tasks until the last minute.*	❍	❍	❍	❍	❍
3.	*When I have to be somewhere, I arrive on time.*	❍	❍	❍	❍	❍
4.	*I keep my desk neat.*	❍	❍	❍	❍	❍
5.	*I like routine.*	❍	❍	❍	❍	❍
6.	*I am never (or rarely) late for class.*	❍	❍	❍	❍	❍
7.	*I like to make schedules.*	❍	❍	❍	❍	❍
8.	*I write tasks down so I won't forget to do them.*	❍	❍	❍	❍	❍
9.	*When I am going somewhere, I am never late.*	❍	❍	❍	❍	❍
10.	*I use computers or cell phones to remind me of deadlines.*	❍	❍	❍	❍	❍
11.	*I know what is in my backpack.*	❍	❍	❍	❍	❍

12. *Each day, I spend a few minutes planning what I am going to do tomorrow.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

13. *I mark dates that are important to me on a calendar.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

14. *People never complain that I am late.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

15. *I like to make lists of things to do.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

16. *My teachers are glad that I'm never late for class.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

17. *I keep my locker neat.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

18. *I'm rarely late for breakfast.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

19. *I finish tests with plenty of time to go over my answers.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

20. *I plan events ahead of time with my friends.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

21. *I have already planned all the things I am going to do tomorrow.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

22. *I am early for practice (sports, music) or anything I might do after school.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

23. *I am one of the first people to get to class.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

24. *I know what I want to do next weekend.*

Never	Rarely	Sometimes	Often	Always
❍	❍	❍	❍	❍

Acknowledgments

The present research was made possible thanks to generous support from Connie Yowell and the John D. and Catherine T. MacArthur Foundation. We also thank Katie Salen for welcoming us to Q2L and for providing support for our efforts. We thank all of the Q2L teachers and designers who understood the importance of our research, accommodated our needs, and gave up some of their precious time for the study. Finally, we thank the students at Q2L for participating in the assessments.

Note

1. Our third assessment session, originally scheduled for September 2010, was delayed for two months because Q2L moved to a new location after the 2009–2010 school year and was still in the process of setting up the technical infrastructure in September and October 2010.

References

Assaraf, O. B.-Z., and N. Orion. 2005. Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching* 42 (5):518–60. doi:10.1002/tea.20061.

Barak, M., and P. Williams. 2007. Learning elemental structures and dynamic processes in technological systems: A cognitive framework. *International Journal of Technology and Design Education* 17 (3):323–40. doi:10.1007/s10798-006-9006-0.

Bransford, J., A. Brown, and R. Cocking. 2000. *How people learn: Brain, mind, and experience and school*. Washington, DC: National Academy Press.

Forrester, J. W. 1994. System dynamics, systems thinking, and soft OR. *System Dynamics Review* 10 (2–3):245–56. doi:10.1002/sdr.4260100211.

Gardner, J. W. 1995. *Self-renewal: The individual and the innovative society*. New York: W. W. Norton.

Gee, J. P. 2005. What would a state of the art instructional video game look like? *Innovate* 1 (6). http://www.innovateonline.info/pdf/vol1_issue6/What_Would_a_State_of_the_Art_Instructional_Video_Game_Look_Like_.pdf (accessed February 22, 2013).

Lave, J., and E. Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511815355.

Liu, O. L., F. Rijmen, C. MacCann, and R. D. Roberts. 2009. Measuring time management abilities for middle school students. *Personality and Individual Differences* 47 (3):174–79. doi:10.1016/j.paid.2009.02.018.

Macan, T. H., C. Shahani, R. L. Dipboye, and A. P. Phillips. 1990. College students' time management: Correlations with academic performance and stress. *Journal of Educational Psychology* 82 (4):760–68. doi:10.1037/0022-0663.82.4.760.

MacCann, C., A. L. Duckworth, and R. D. Roberts. 2009. Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences* 19 (4):451–58. doi:10.1016/j.lindif.2009.03.007.

Ossimitz, G. 2000. The development of systems thinking skills using system dynamics modeling tools. http://wwwu.uni-klu.ac.at/gossimit/sdyn/gdm_eng.htm (accessed May 10, 2010).

Partnership for 21st Century Skills. 2006. *Assessment of 21st century skills*. Washington, DC: Partnership for 21st Century Skills.

Resnick, M. 2007. Sowing the seeds for a more creative society. *Learning and Leading with Technology* 35 (4):18–22.

Roberts, R. D., R. Schulze, and J. Minsky. 2006. The relation of time management dimensions to scholastic outcomes. Presentation at Annual Meeting of the American Educational Research Association, San Francisco.

Rysavy, D. M., and G. C. Sales. 1991. Cooperative learning in computer-based instruction. *Educational Technology Research and Development* 39 (2):70–79. doi:10.1007/BF02298155.

Salen, K., R. Torres, L. Wolozin, R. Rufo-Tepper, and A. Shapiro. 2011. *Quest to learn: Developing the school for digital kids*. Cambridge, MA: MIT Press.

Salisbury, D. F. 1996. *Five technologies for educational change: Systems thinking, systems design, quality science, change management*. Englewood Cliffs, NJ: Educational Technology Publications.

Shute, V. J., and R. Torres. 2012. Where streams converge: Using evidence-centered design to assess Quest to Learn. In *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research*, ed. M. Mayrath, J. Clarke-Midura, and D. H. Robinson, 91–124. Charlotte, NC: Information Age Publishing.

Sweeney, L. B., and J. D. Sterman. 2007. Thinking about systems: Student and teacher conceptions of natural and social systems. *System Dynamics Review* 23 (2/3):285–312. doi:10.1002/sdr.366.

Tindale, R. S., S. Stawiski, and E. Jacobs. 2008. Shared cognition and group learning. In *Work group learning: Understanding, improving and assessing how groups learn in organizations*, ed. V. I. Sessa and M. London, 73–90. New York: Lawrence Erlbaum Associates.

Torres, R. 2009. Learning on a 21st century platform: Gamestar Mechanic as a means to game design and systems-thinking skills within a nodal ecology. PhD diss., New York University.

Totten, S., T. Sills, A. Digby, and P. Russ. 1991. *Cooperative learning: A guide to research*. New York: Garland.

Van den Bossche, P., M. Segers, and P. A. Kirschner. 2006. Social and cognitive factors driving teamwork in collaborative learning environments. *Small Group Research* 37 (5):490–521. doi:10.1177/1046496406292938.

Vygotsky, L. S. 1978. *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wang, L., C. MacCann, X. Zhuang, O. L. Liu, and R. D. Roberts. 2009. Assessing teamwork skills: A multi-method approach. *Canadian Journal of School Psychology* 24 (2):108–24. doi:10.1177/0829573509335470.

Zhuang, X., C. MacCann, L. Wang, O. L. Liu, and R. D. Roberts. 2008. *Development and validity evidence supporting a teamwork and collaboration assessment for high school students*. ETS Research Report, RR-08-50. Princeton, NJ.