A VOLUME IN THE MARCES BOOK SERIES

TEST FAIRNESS IN THE NEW GENERATION OF LARGE-SCALE ASSESSMENT

Hong Jiao &
Robert W. Lissitz, Editors

Test Fairness in the New Generation of Large-Scale Assessment

A Volume in
The MARCES Book Series

Series Editors

Hong Jiao and Robert W. Lissitz
University of Maryland

The MARCES Book Series

Hong Jiao and Robert W. Lissitz, Series Editors

Test Fairness in the New Generation of Large-Scale Assessment (2017)
edited by Hong Jiao and Robert W. Lissitz

The Next Generation of Testing: Common Core Standards, Smarter-Balanced, PARCC, and the Nationwide Testing Movement (2015)
edited by Hong Jiao and Robert W. Lissitz

Value Added Modeling and Growth Modeling With Particular Application to Teacher and School Effectiveness (2015)
edited by Robert W. Lissitz and Hong Jiao

Informing the Practice of Teaching Using Formative and Interim Assessment: A Systems Approach (2013)
edited by Robert W. Lissitz

Computers and Their Impact on State Assessments: Recent History and Predictions for the Future (2012)
edited by Robert W. Lissitz and Hong Jiao

The Concept of Validity: Revisions, New Directions and Applications (2009)
edited by Robert W. Lissitz

Test Fairness in the New Generation of Large-Scale Assessment

edited by

Hong Jiao

and

Robert W. Lissitz
University of Maryland

IAP

INFORMATION AGE PUBLISHING, INC.
Charlotte, NC • www.infoagepub.com

Library of Congress Cataloging-in-Publication Data

CIP record for this book is available from the Library of Congress
http://www.loc.gov

ISBNs: 978-1-68123-893-7 (Paperback)

978-1-68123-894-4 (Hardcover)

978-1-68123-895-1 (ebook)

Printed in the United States of America

CONTENTS

CHAPTER 1

RESOLVING THE PARADOX OF RICH PERFORMANCE TASKS

Robert J. Mislevy

ABSTRACT

Interest in rich performance tasks has been increasing, due in part to advances in learning science that show their value in learning and in part to advances in technology that allay many issues of cost, scale, and quality. To understand the value of performance tasks as an assessment method requires ideas from learning science and evidentiary reasoning as well as from measurement. This chapter uses these ideas to explore the implications of adding depth, context, and interactivity to tasks, as they might be used in a variety of situations and for various purposes. It shows how inference can be strengthened within contexts and substantive contents, which is particularly well-suited to assessment integrated with learning. However, the same contextualization can contribute construct-irrelevant variance for inference for broader inferences and to other contexts and substantive content. The performance expectations of the Next Generation Science Standards and a game-based assessment called SimCityEDU: Pollution Challenge! for developing systems thinking are used to illustrate ideas.

Test Fairness in the New Generation of Large-Scale Assessment, pp. 1–46

1.0 INTRODUCTION

1.1 Performance Tasks In Educational Assessment

Advances in technology and learning science are transforming the world of education, and with it the world of assessment. This chapter notes some key advances—in particular, a sociocognitive perspective on learning and digital environments that enable performance assessment to be scaled up efficiently—and examines their implications for the roles that performance assessment can play in that new world. The chapter draws on developments from measurement and assessment design that help us understand when and how to use performance assessments effectively.

Performance assessment is not new. Medieval apprentices produced masterpieces to demonstrate they had the necessary skills to enter a guild. The first edition of *Educational Measurement* (Lindquist, 1951) included a chapter by Ryans and Frederiksen (1951) on the topic. It focused on industrial and professional applications. The standards movement of the 1980s and 1990s saw more widespread application of performance assessment in large-scale testing, argued to be better evidence for educative goals (Resnick, 1994). Their use declined due to relatively high costs and the generalizability issues that are one of the issues that discussed here.

A number of factors have come together to spur renewed interest in performance assessment. One key development is the capability to produce interactive computer environments at large scale, to capture and analyze voluminous data from those environments, and to evaluate performances automatically. Tasks that could be done only on a small scale at high cost, for example, can be accomplished by digital means at a fraction of the cost, and can be administered virtually anywhere, anytime. Another development is a broader conception of learning, beyond forms of knowledge and skill that can be easily assessed with simple tasks. For example, the Next Generation Science Standards (NGSS; National Research Council, 2012) offers “performance expectations” as representative sketches of rich tasks that integrate disciplinary ideas, science and engineering practices, and cross-cutting themes such as “systems and system models.” Performance assessment is energizing discussion across all levels of education and across disciplines. It is central to standards movements and to new forms of instruction in both schools and online learning. It has spawned new products, new industries, and new job titles. And the issues addressed here lie within every instance of its application.

There is no precise definition of “rich performance tasks,” but there are family resemblances among tasks that most observers would agree merit the term, and clear contrasts with tasks that do not. Rich performance tasks usually have some or all of the following features: Interactivity, multiple steps, openness and construction in responding, contextualization of the task, require extended amounts of time, integration of multiple aspects of knowledge and skill, and requirements for some higher-level skills such as critical thinking, problem-solving, systems thinking, communication, and collaboration. Some are designed to resemble domain-specific activities that are required of professionals in a domain, such as performing a laboratory experiment or trouble-shooting a computer network. They contrast with familiar tests that typically use choice-based responses, have little context, provide for minimal interaction, and do not evaluate the processes that constitute performance. The running example will be presented in more depth shortly, but two quick examples illustrate the idea:

- The National Board of Medical Examiners (NBME) evaluates unique paths of actions in the Primum computer-simulated patient management problems (Dillon & Clauser, 2009). Medical licensure candidates evaluate patients, decide what treatments to employ, monitor progress, and adjust treatments in accordance with the patient's response.
- In 2015, the Program for International Student Assessment (PISA) assessed collaborative problem-solving competencies (Organization for Economic Co-operation and Development, 2013). Conversational agents represented peers with a range of skills and abilities and other characteristics, as well as behavior—team members who initiate ideas and support and praise others versus team members who interrupt and criticize others and propose misleading strategies (Davey, Ferrara, Holland, Shavelson, Webb, & Wise, 2015). A test taker might collaborate with a computer agent to determine the best water and other conditions for fish in an aquarium.

1.2 The Paradox of Performance Tasks

The Communication Within the Curriculum Speaking Centers (CWIC) at the University of Pennsylvania provides support for teachers who are planning debates for their students.[1] The most important aspect of any debate, they advise, is the topic, a statement that people could either affirm or negate. "Ideally people will be able to affirm or negate the resolution for a variety of reasons, with many possibilities for constructing sophisticated positions on each side." Robert Branham (2013) proposed that true debate depends on the presence of four characteristics: "*Development*, through which arguments are advanced and supported; *Clash*, through which arguments are properly disputed; *Extension*, through which arguments are defended against refutation; and *Perspective*, through which individual arguments are related to the larger question at hand" (p. 22, original emphasis). With these characteristics of debate in mind, I suggest that the following statement would be a good debate topic:

> **RESOLVED:** *Rich performance assessments make for improved assessment practice.*

To begin, the literature offers compelling, well-researched arguments for both the affirmative and negative teams (Davey et al., 2015). Its proponents advance several benefits of performance assessment. Performance assessment generates observable performance of higher-order thinking with generic and/or domain-specific content in contexts, they argue. It adds value in both the types of complex skills that can be assessed and the types of instructional strategies it reinforces and informs. It signals the importance of both higher-order thinking and applying such thinking to accomplish goals in real-world contexts.

On the other hand, evidence from large-scale performance assessments going back to the 1990s repeatedly advises caution in using performance assessment (Linn, 2000). Studies reveal poor generalizability across raters, across time, across tasks, and across occasions.[2] There are potential effects of lack of opportunity to learn. There can be effects of construct-irrelevant requirements with respect to language, expectations, materials, evaluation methods, and so on. (Research on these effects in digital environments is still in early stages; Clarke-Midura & Dede, 2010. We will see how some key evidentiary issues that contributed to the previous results arise with new forms of performance assessment.)

Both the affirmative and negative cases make valid points. This chapter extends from these clashing positions, guided by our growing understanding of the cognitive and social interplay of human learning and acting. We will see that a resolution requires several elements: The contextualization of the task with respect to the students' instruction; the target of inference; the degree to which the target inference is connected to the students' instruction; the relationship of the task to the students' past experience and learning; and, finally, what the assessment user knows or does not know about these relationships. Recurring configurations of these factors can be described as assessment use cases (Gorin & Mislevy, 2013).

So, do rich performance assessments make for improved assessment practice? In certain assessment use cases, the answer is a resounding yes; in others, an emphatic no.

1.3 Roadmap of the Chapter

The remainder of the chapter develops the perspective behind the resolution—just why, through the lenses of measurement principles and sociocognitive psychology, the properties of rich performance assessments make for better assessment practice in some use cases and worse in others. Observations will be made along the way concerning validity, generalizability, and fairness.

Section 2 presents a running example to help ground the discussion, a game-based simulation task called *SimCityEDU: Pollution Challenge* (Mislevy, Corrigan, Oranje, DiCerbo, John, Bauer, Hoffman, von Davier, & Hao, 2014). Section 3 gives additional background for the NGSS, a currently important framework that advocates rich performance tasks.

Section 4 sketches a sociocognitive perspective on learning and performance, and notes implications for situated action, learning, and assessment that bear on the utility of performance assessment. Sections 5 and 6 discuss the implications in greater detail, focusing respectively on key cognitive and social aspects. The nature of higher-level skills and the role of students' backgrounds receive special attention.

Section 7 reviews assessment interpretation arguments, highlighting the strands that are central to the discussion. Section 8 describes four familiar assessment use cases, chosen to bring out different evidentiary properties of performance tasks.

Section 9 is where the ideas developed in the preceding sections finally come together. It discusses the implications of using rich performance tasks in each of the four exemplar use cases. We see the ones in which rich performance assessments is particularly attractive and others in which it is not, and discuss why this is so.

The major resolution having been completed, Section 10 adds some practical notes on strategies for using performance assessments effectively. Section 11 summarizes the main conclusions.

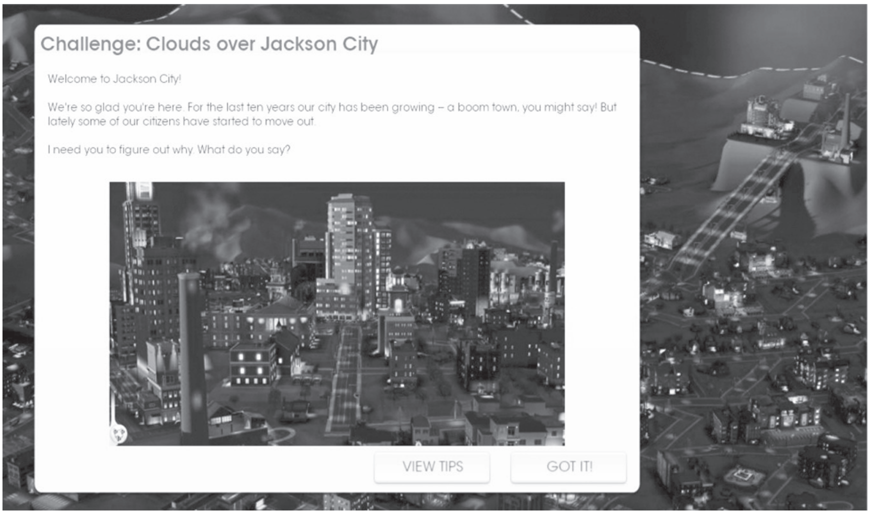
2.0 SIMCITYEDU: POLLUTION CHALLENGE

The Jackson City scenario in GlassLab's SimCityEDU game-based assessment (http://www.playfully.org/games/SC) will serve as a running example of a rich performance task. Based on the SimCity commercial game, SimCityEDU presents a series of challenges in which players tackle a city's problems in ways that require balancing environmental impact, infrastructure needs, and employment. The game scenarios help players learn about systems thinking, with formative assessment integrated into play. Systems thinking is a cross-cutting concept in the Next Generation Science Standards (NGSS; NGSS Lead States, 2013a). It is an understanding of how components of a system influence each other, incorporating concepts such as feedback, adaptation, emergent behavior, and unintended consequences. The assessment is built on a learning progression, or a framework for the development of student understanding in this area. The game's challenges reflect the levels in the learning progression for systems thinking shown in Table 1.1 (from Mislevy et al., 2014).

Table 1.1. The Systems-Thinking Learning Progression From SimCityEDU

Level	*Competency Level Description*
5	Students have a ***globally coherent understanding*** of many aspects of systems thinking in many contexts. They can analyze of moderately complex system that includes multiple variables, including several hidden variables, feedback spread out in space and time, and emergent behaviors that requires understanding a system at multiple levels, with multiple causes interacting to create complex emergent effects (corresponding to level 5 in Brown, 2005).
4	Students can relate ***multiple*** causes to multiple effects as long as they behave in simple ruleful ways (e.g., cases in which all causes are needed for the effect to occur, cases in which all causes contribute independently to the amount of the effect as in Jackson City, etc.; .i.e., the causes are not emergent but are instead explainable in terms of the causal component parts. This level is consistent with Brown's (2005) conceptual depth level 4. Students can apply this scope of understanding within a wider range of contexts than in prior levels.
3	Students have a ***locally coherent*** understanding of many aspects of systems. Students can use system thinking terms to describe components and system relations in some contexts and use different representations. They can use models to represent bivariate cause and effect relations along with strong justifications. They can relate binary combinations of hidden and directly observable combinations, and even single causes to multiple effects. I.e. they are less prone to common misconceptions but still are limited linear thinking with single causes (which may or may not be chained together.) They have a rudimentary understanding of negative feedback and can use it to explain and predict change in behavior of a system over time. They still are not able to consistently understand and analyze a system at different levels (Cheng, Ructtinger, Fujii, & Mislevy, 2010).
2	Students have an ***elemental*** understanding (Brown, 2005, p. 7) of some aspects of systems—they can use models to represent simple, single cause and effect relations but without strong justification i.e. they are still prone to common misconceptions, e.g., they tend to only relate macrolevel, directly observable causes and effects rather than identifying hidden variables and factors. This is due in part to not being able to understand and analyze a system at different levels (Cheng et al., 2010).They are better at explaining than predicting.
1	Students have a ***fragmented*** understanding of aspects of systems. They may have partial knowledge of some of the definitions of system terms but cannot use them in a consistent nor strongly coherent manner. While they can identify outcome variables (e.g., stocks that are explicitly part of the goal state), they are not able to track a causal link and they largely focus on macro-level directly-observable variables. Their predictions and explanations are acausal, more assertions than cause and effect relations (e.g. "things happen because that's the way they are" Brown, 2005, p. 7).

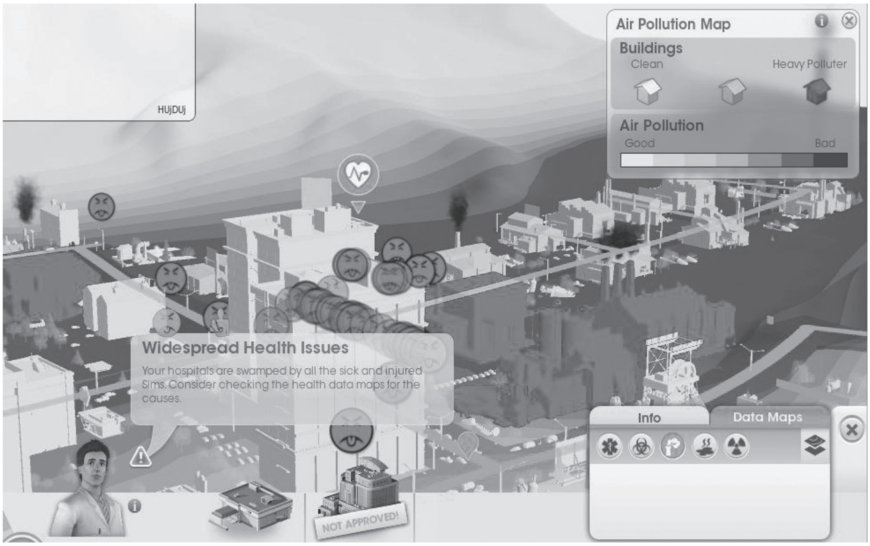
In the Jackson City challenge, the player (in the role of mayor) enters the city (Figure 1.1) and is told that residents seem unhappy and are leaving. Interaction with the Sim characters reveals that they are having trouble with air pollution. Players can explore data maps that show which buildings are polluting (Figure 1.2), how power is dispersed in the city, and how various areas are zoned. They discover that coal plants are the biggest cause of pollution in the city. However, coal plants also provide much of the power in the city. Power impacts both resident happiness and jobs (unpowered businesses shut down).



Source: Mislevy et al. (2014) (used with permission from the Institute of Play).

Figure 1.1. Initial view of Jackson City.

In the game, players can bulldoze buildings, place new power structures (wind, solar, or coal generated), build new roads to expand their city, and zone and dezone residential, commercial, and industrial areas in order to achieve their goals. They can monitor the effects of their actions on pollution and jobs with on-screen thermometers. The player experience is one of tackling a troubleshooting challenge; yet at the same time, players' actions are captured and provide evidence for their level of systems thinking. For example, a player might focus solely on the relationship between the coal plants and pollution, and bulldoze coal plants. This action is consistent with Level 2 in the learning progression. A player may recognize the multiple effects of coal plants, both causing pollution and providing power. This player would be observed placing alternative energy options and bulldozing coal plants, but taking no actions that suggest attention

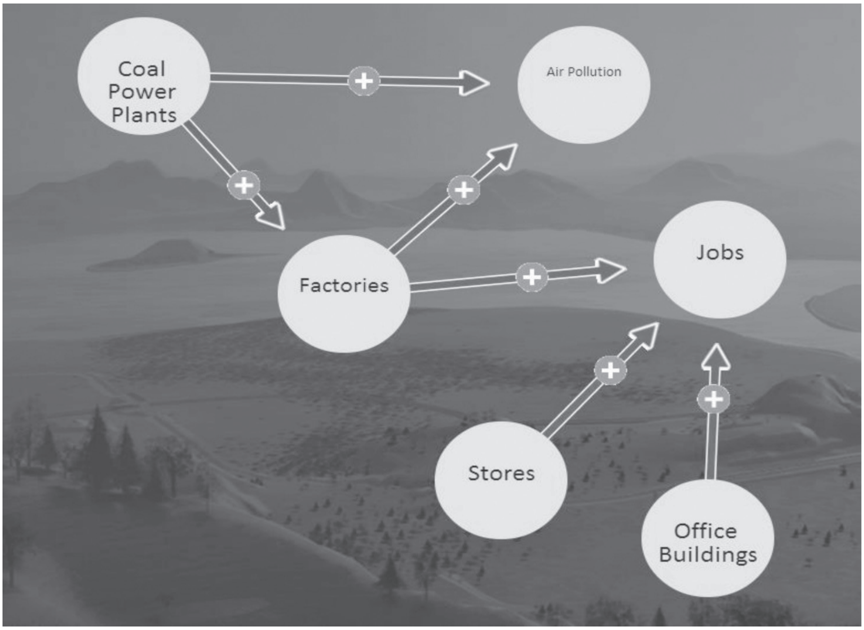


Source: Mislevy et al. (2014) (used with permission from the Institute of Play).

Figure 1.2. Use of a tool to monitor amounts and locations of pollution production.

to the unemployment problem. This is consistent with Level 3. Actions consistent with Level 4 thinking would address the pollution and power tradeoffs and also create new commercial zones to help increase available jobs. These actions and sequences are extracted from log files and provide evidence in a Bayesian network measurement model (DiCerbo et al., 2015; DiCerbo, Mislevy, & Behrens, 2016; Mislevy et al., 2014). The outcome is a posterior probability distribution across the levels that the player seems to be thinking at, given her several actions throughout her solution.

The instructional support that GlassLab developed for using SimCityEDU in a classroom plays an important role in students' learning and in the evidentiary value of their play as assessment information. GlassLab did not plan for learning to come from play alone. Students' in-game play is interspersed with guided discussions about systems concepts and representations, and how they relate to what is happening in Jackson City. Figure 1.3, for example, is a system diagram tool students use to help them understand what is happening in one challenge. The students also complete these diagrams before and after a challenge as pre-designated assessment information. The students themselves receive feedback individually, and the teacher receives summary reports on the class in order to help guide discussions.



Source: Mislevy et al. (2014) (used with permission from the Institute of Play).

Figure 1.3. Example of a Jackson City system diagram.

3.0 NGSS PERFORMANCE EXPECTATIONS

The Next Generation Science Standards (National Research Council, 2012; NGSS Lead States, 2013a) presents a framework and standards for instruction and assessment. The NGSS are meant to reflect the inherent complexity in scientific understanding and reasoning as it exists in the real world. They address not only core disciplinary ideas, but also scientific practices such as *developing and using models* and *planning and carrying out investigations* and cross-cutting concepts such as *systems and system models* and *structure and function*. Compared to previous science standards, the NGSS enacts several conceptual shifts:

- K–12 science education should reflect real world interconnections.
- All science practices and crosscutting concepts are used in teaching all core ideas.
- Science concepts build coherently across K–12.
- The NGSS focuses on deeper understanding and application of content.

- Science instruction and assessment should coordinate with English Language Arts and Mathematics standards.

The NGSS architecture intentionally gives considerable latitude for instructional and assessment design choices. To support educators, it provides *performance expectations* to operationally define the standards. Performance expectations are the assessable statements of what students should know and be able to do, and are written to combine the disciplinary idea, practice, and cross-cutting concept dimensions. While they provide descriptions of the achievements students should be able to demonstrate at grade-level bands, they do not translate directly into any single instructional activity or assessment task. Performance expectations are meant "to communicate a 'big idea' that combines content from the three foundation boxes" (NGSS Lead States., 2013a, p. 2).

The NGSS authors want students emerge from science and engineering education with competency in the key practices and concepts as they interact with core disciplinary ideas. But designing instructional and assessment activities to reflect real-world such problem-solving requires specific contexts, formats, and materials. To help designers make decisions about specific instructional and assessment tasks, the NGSS includes clarification statements for many of the performance expectations provide some guidance as to some of the contexts in which one might develop activities. These statements highlight the fact that there are a variety of contexts, each with its own context-specific content knowledge, in which one might choose to teach or assess the same expectation.

For example, Table 1.2 shows *4-ESS3-1 Earth and Human Activity* (NGSS Lead States, 2013b). The Jackson City scenario can be considered one of many possible instantiations of this performance expectation. Substituting *systems and system models* for *cause and effect* would make the fit even better. We will return repeatedly to the point that assessing systems thinking with the rich tasks that NGSS advocates necessarily involves some particular practice(s), some particular system(s), in some particular context(s).

4.0 A SOCIOCOGNITIVE PERSPECTIVE

4.1 The Basic Idea

Educational assessment evolved under trait and behavioral psychology. To design and use more complex assessments—interactive, integrated, and constructive, like Jackson City—requires a perspective that can address the moment-by-moment nature of how people act and learn, and the ocean of social and cultural patterns that give meaning to that acting and learning. Casting the term broadly, this a situative, sociocognitive perspective. It encompasses findings that connect many strands of cognitive and social research, and can be argued to encompass insights from the trait, behavioral, and information-processing perspectives (Greeno, Collins, & Resnick, 1997). This section is a brief sketch of such a perspective, highlighting ideas that are key to performance tasks.

Table 1.2. A Performance Expectation From the Next Generation Science Standard

4-ESS3-1 Earth and Human Activity

Students who demonstrate understanding can:

> Obtain and combine information to describe that energy and fuels are derived from natural resources and their uses affect the environment. [Clarification Statement: Examples of renewable energy resources could include wind energy, water behind dams, and sunlight; nonrenewable energy resources are fossil fuels and fissile materials. Examples of environmental effects could include loss of habitat due to dams, loss of habitat due to surface mining, and air pollution from burning of fossil fuels.]

The performance expectation above was developed using the following elements from the NRC document *A Framework for K–12 Science Education*:

Science and Engineering Practices	**Disciplinary Core Ideas**	**Crosscutting Concepts**
Obtaining, Evaluating, and Communicating Information • Obtain and combine information from books and other reliable media to explain phenomena.	ESS3.A: Natural Resources • Energy and fuels that humans use are derived from natural sources, and their use affects the environment in multiple ways. Some resources are renewable over time, and others are not.	Cause and Effect • Cause and effect relationships are routinely identified and used to explain change. - - - - - - - - - - - - - - - - **Connections to Engineering, Technology, and Applications of Science** Interdependence of Science, Engineering, and Technology • Knowledge of relevant scientific concepts and research findings is important in engineering. Influence of Engineering, Technology, and Science on Society and the Natural World • Over time, people's needs and wants change, as do their demands for new and improved technologies.

The "socio-" in "sociocognitive" highlights the patterns of knowledge and activity that structure the interactions people have with the world and other people. These include the structures and ways of using language, knowledge representations, and cultural models, and of the patterns of activities of families, communities, personal interactions, and classrooms and workplaces (Wertsch, 1994). Collectively we may call them linguistic, cultural, and substantive (LCS) patterns. Of particular interest for present purposes are the kinds of things we learn for school and work, such as the core disciplinary ideas, practices, and cross-cutting concepts in the NGSS.

The "-cognitive" highlights within-person cognitive patterns, from large to small and across different levels. These are traces of each individual's past experiences, continually assembled, adapted, and revised to make meanings and guide actions in each new situation. Young (2009) and Hammer, Elby, Scherr, and Redish (2005) use the term "resources" to describe unique within-person patterns of relationships among knowledge, relationships, actions, feelings, and motives we develop and assemble to make our way through the physical and social world.

A sociocognitive perspective addresses the interplay among these levels: Cognitive processes within individuals give rise to their actions in the human-level activities we experience, as we negotiate the situations which, while unique in their particulars, build simultaneously around LCS patterns at many levels. Researchers from both cognitive and social bents have used an iceberg metaphor to emphasize how little we are aware of consciously as we activate and assemble numberless cognitive resources to recognize, interact with, and create the ever-changing flux of situations structured around numberless LCS patterns (e.g., Fauconnier, 1999; Haggard, 2005).

4.2 Situations, Actions, and Resources

Several confluences must occur between patterns in a situation and patterns in an individual for the familiar activities that comprise everyday life, from buying groceries, to planning a trip with a friend, to solving Jackson City's pollution problem. In Jackson City, for example, the situation at a particular moment of play is structured jointly on myriad LCS patterns, of various kinds and at many grainsizes. A player Sally must correspondingly draw on resources she has developed to make sense of the unfolding situation, and figure out what to do next. She is blending LCS patterns that the particulars of the immediate situation have activated (Fauconnier & Turner, 2002; Kintsch, 1998)—continually acting, revising, and all the while, building resources. She must understand something about mayors, cities, jobs, and power plants. She must understand English well enough to make sense of help, scenario descriptions, and simulated citizens' complaints. She must navigate in a SimCity world, moving from one view to another, and do things like zoom, plop, and hover. She must coordinate her play and understanding of Jackson City with all of the activity patterns and knowledge patterns of the classroom, particularly the ones that create the broader instructional frame that envelops her actions in Jackson City.

The designers of SimCityEDU hope that Sally will develop resources from this experience that are useful beyond SimCityEDU—that are useful for thinking about other situations Sally might encounter that can productively be understood through these system concepts. They hope that the resources have been developed such that the features of these other unique situations will nevertheless activate these more general "systems" resources. Sally comes to SimCityEDU with a network of understanding of the words "cause" and "effect," for example, built up from her experiences with these words at home and school, with friends and family, in books and television, and so on. Her understanding of these words overlaps some with the more technical ways scientists use the same words—shared definitions and representations, and the attributes and phenomena they associate with the words from their own unique experiences. The goal is that interacting with Jackson City's jobs-and-pollution system and using the more scientific terms and diagrams in this crafted environment, Sally will experience some of the patterns the words are used for in science, and expand her semantic networks in ways that begin to overlap more with those meanings (Roth, 2009).

Kintsch and Greeno (1985) suggested how solving science problems involves constructing a blend of abstracted disciplinary models, linguistic structures that communicate relationships among the models and real-world phenomena, and the particulars of the unique situation at hand. This kind of generalization does not happen automatically, for resources are initially tied closely to the conditions of learning (Greeno, 1998). Over time, and with more experiences that are variations across LCS themes, sometimes resources will be developed that are more abstract and activated more widely. This is the case for many of the proficiencies we develop as readers. It is not necessarily the case for the problems we learn to solve at the end of the chapters of a physics text. And we may develop resources as research chemists, say, to communicate quite effectively to other research chemists; but employing the same resources in what we misperceive to be the same way could prove disastrous on the witness stand.

5.0 IMPLICATIONS FOR LEARNING THAT HIGHLIGHT A COGNITIVE ASPECT

This section looks more closely at results for a particularly cognitive aspect of learning, namely patterns in how an individual's resources develop. Simplified topographical maps suggest the way resources for the kinds of learning the NGSS promotes can occur (Hammer et al., 2005; Young & He, 1998). Implications for the meanings of learning progressions and higher-level skills are noted.

5.1 Topographical Maps

We learn from experience in unique situations structured around LCS patterns at many levels, and the resources we develop are initially tied to the circumstances of learning. An individual's trajectory of experience cultivates clusters of resources and dense interconnections with regard to topics and practices that occupy their interests and activities. This is obvious for adults in their occupations and hobbies, but we also see it in young children who often become quite interested in some area—"islands of expertise," Crowley and Jacobs (2002) call them. They described a child who received a *Thomas the Tank Engine* book on his second birthday. Over the next year he learned as much as he could first about Thomas, then about trains more generally including rather technical information, all supported by his parents in conversations, visits to museums, make-believe games, and so on. With his deep knowledge in this particular area, he could carry out more sophisticated reasoning and explanations than he could in other areas. For example, his mother helped him understand a boiling tea kettle by drawing connections to how steam engines work.

No less than children, we are all characterized by the islands of expertise we develop in our own trajectories through situations in the cultures and the affinity groups we move in. We develop more islands over time, build connections across them, and in some cases develop resources for more general schemas that could be applied[3] to new situations—cross-cutting concepts, as it were.

Figure 1.4 suggests these processes. For the sake of illustration, imagine a science curriculum that uses learning experiences built around NGSS performance expectations. Working through SimCityEDU in class would be a middle-school example. Panel (a) represents a student Carlos at the very beginning of the curriculum, before these structured experiences. Suppose the X and Y axes correspond to core disciplinary ideas and cross-cutting themes, and the height Z corresponds to proficiency, in terms of resources Carlos can bring to bear on a situation he might encounter. This is a ridiculously simple picture, not only because each dimension would have vastly more possible topics and themes, but also because there would be many more dimensions that would concern practices, contexts, materials, language, mathematical models and practices, and so on. Nevertheless, the point is that Carlos enters the picture with quite modest resources, but they are stronger with respect to some idea-by-theme combinations and sparser in others as they developed in his previous experiences.

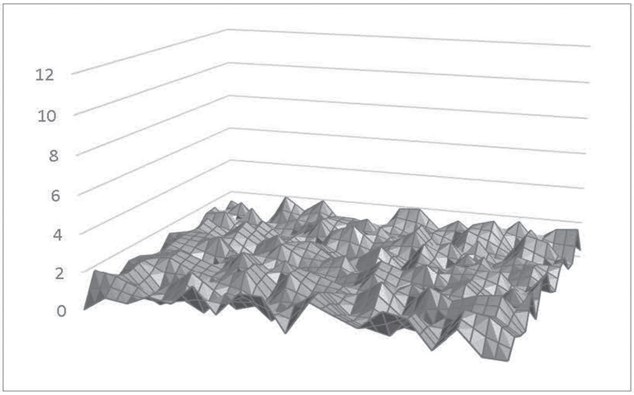
Panel (b) is the result after Carlos has worked through two in-depth investigations. We first notice spikes where resources have developed around the particular combinations addressed in these tasks with regard to core ideas and cross-cutting themes (as well as practices, representations, and so forth, on the hundred other dimensions). There are peaks for the foci of the tasks in the formative assessment Carlos and his teacher received feedback from as he worked through the investigation.

Note that tasks in this neighborhood are hard in one sense, but just right in a different sense. They are hard *marginally*, in that few fourth graders sampled randomly across the nation would have the particular combination of experiences involving the systems thinking representations, the jobs-and-pollution system, and the familiarity with the simulation environment of this SimCityEDU's Jackson City challenge. But *conditionally*, they are just right to provide information about Carlos, given his experiences so far in the classroom discussions and the series of SimCityEDU challenges he has worked through so far. These are very particular experiences that help locate Carlos's zone of proximal development, to use Vygotsky's psychological term; and at the same time, a region of maximum information, to use a term from measurement (Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012).

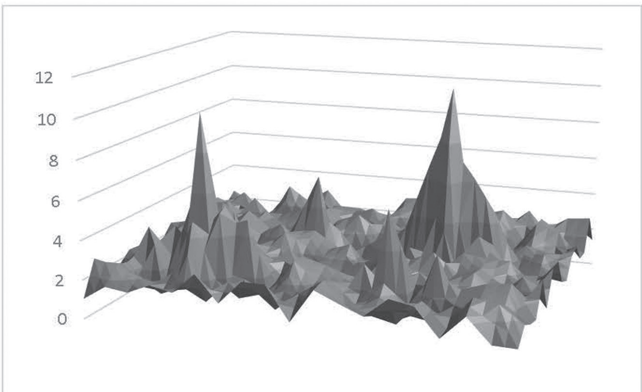
In addition to the peaks themselves, we also note ridges along the dimensions that were addressed. These represent resources that have developed in the experience that might have "hooks" that could be activated in some other contexts or with other disciplinary ideas. Carlos might encounter a new situation, say the relationship between the populations of wolves and moose on Isle Royale, and be moved to think about them in terms of the systems concepts he worked with in Jackson City. We notice too that the surface is a bit higher on average. This represents how increased resources, spotty as they are and unpredictable in their activation as they may be, have increased Carlos's capabilities to make sense of a new situation he encounters, to recognize important features in terms of more general LCS patterns, to have choices for acting, and to be able to create new resources and connect them with current ones.

Panel (c) looks again at Carlos after a succession of such experiences, involving various combinations of disciplinary ideas and cross-cutting themes. There are still peaks and valleys, but there more peaks and more ridges that bridge valleys. The overall surface is higher still.

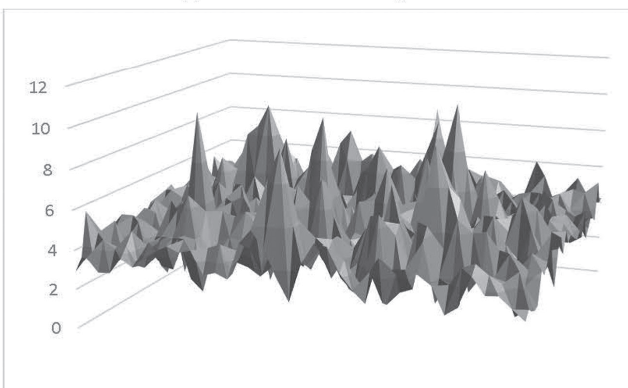
One point that will become important with regard to measurement is that Carlos's peaks and valleys are not in the same places as other students'. Some are similar; if Carlos and his classmates have all worked through SimCityEDU, they will have similarities in those regions where they have



(a) Before instruction.



(b) After two rich learning tasks.



(c) After many rich learning tasks.

Figure 1.4. Hypothetical simplified topography of proficiency.

shared in-depth experiences, in particular ways of thinking about systems and especially jobs-and-pollution systems. But if Carlos is a Thomas the Tank aficionado he may well have more of a propensity to think through a locomotive repair problem in systems terms than Sally. Conversely, Sally, who is growing up on a farm, will be more apt to explain the relationship between bees and crop yields using systems concepts. These kinds of effects contribute to person-by-task variance in generalizability analyses when the target inference is analogous to the average height of these topographs.

5.2 Instructional Strategies

The instructional challenge is how to structure students' experiences to best build bridges and increase the overall height of the surface. In traditional psychological terms, this is the problem of transfer. In sociocognitive terms, it is developing resources that can be activated further beyond initial conditions of learning (Hammer et al., 2005), and that are more likely to capitalize on opportunities for future learning (Bransford & Schwartz, 1999). Educators and learning scientists and researchers have advanced a number of strategies to this end. The ones mentioned below are powerful for designing instruction. They have powerful, sometimes subtle, counterparts for assessment.

Three approaches greatly help a student develop broadly applicable resources, such as being able to put NGSS's disciplinary ideas to practical use in different contexts, to gain insights by seeing situations in terms of cross-cutting themes, or to carry out scientific practices in new situations. First, the recognition that such resources begin developing in particular contexts—tangible, actionable, contexts, where a student uses them to interact with some situation in the world—to solve a problem, to investigate a phenomenon, to explain a solution to someone else. Second, it is especially powerful when those concrete experiences leverage the knowledge students bring to the situation, as in the Thomas the Tank example, and in science learning that starts with everyday experiences, and in analyses of literary devices as they are used in familiar ways of using language in families and communities practices (Lee, 2008). Third, it takes multiple contexts that vary in particulars but are similar with respect to the higher-level ideas, such as the Jackson City pollution system and the wolves-and-moose food web and population system. As James Gee has put it, "Abstract representations of knowledge, if they exist at all, reside at the end of long chains of situated activity."[4] This insight led to the NGSS recommendations for reflecting real-world connections and integrating disciplinary ideas, practices, and cross-cutting themes.

But having such experiences alone does not necessarily produce the higher level resources. One can become adept at problem-solving in the challenging video game Halo, but not improve at all at how one might solve problems in troubleshooting trucks, managing employees, or herding sheep. An effective strategy is to explicitly connect the situated experiences with the abstracted concepts and representational forms. This insight led to SimCityEDU's designers to embed play in a larger conversation using the vocabulary and representations of systems.

Recall that acting in any real-world situation involves many kinds of LCS patterns at many levels, even answering the simplest multiple-choice test item—indeed, even knowing what a test is, what this genre "multiple-choice item" means, or the expectation that you should answer it and the affordances you have to do so. In an instruction or assessment situation, any of the LCS patterns it explicitly draws on, or many more that are unknowingly presumed, can stymie a student if she lacks some necessary but construct-irrelevant resources, or activates some otherwise effective resources that do not match the situation's expectations. Section 9 will address this issue as a potential source of invalidity and unfairness. For instruction, it means that for some students, what might appear to be an opportunity to learn is actually not (Moss, Pullin, Haertel, Gee, & Young, 2008). Given that rich learning/assessment tasks like Jackson City, which integrate disciplinary knowledge and higher-level schemas in a grounded active context, hold value for learning, how can we avoid derailing the exercise by mismatching LCS demands and students' resources?

One effective instructional strategy is creating rich experiences which do indeed integrate a variety of contextual and substantive LCS patterns with learning targets such as core ideas, practices, and themes—yet which are matched to students so that we know they have already developed many of the resources that are needed along with the targeted ones. One way to implement this strategy is to design a sequence of tasks that spirals to increasing levels of proficiency on certain dimensions, while keeping others within familiar regions (Robinson, 2010; Songer, Kelcey, & Gotwals, 2009). Another is to adapt task schemas to what is known about students (Liu & Haertel, 2011)—an investigation of the effects of natural forces on terrain, for example, fleshed out in the context of local terrain and forces. Some critical elements of knowledge will thus be familiar to each student and not impede their work with the targeted learning objectives, even though "the task" would be different for students in different locales. These strategies can also be understood as reducing extraneous cognitive load (Sweller, Van Merrienboer, & Paas, 1998). In terms of Figure 1.4, a sequence of tasks could be imagined as building peaks at different locations but along ridges defined by the targeted LCS patterns.