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

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

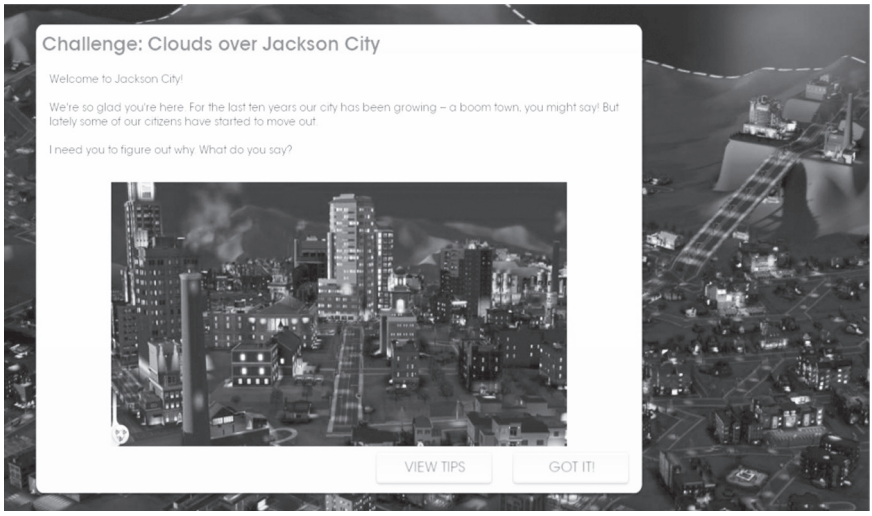
The Jackson City scenario in GlassLab's SimCityEDU game-based assessment (http://www.playfully.org/games/SC) will serve as a running example of a rich performance task. Based on the SimCity commercial game, SimCityEDU presents a series of challenges in which players tackle a city's problems in ways that 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).

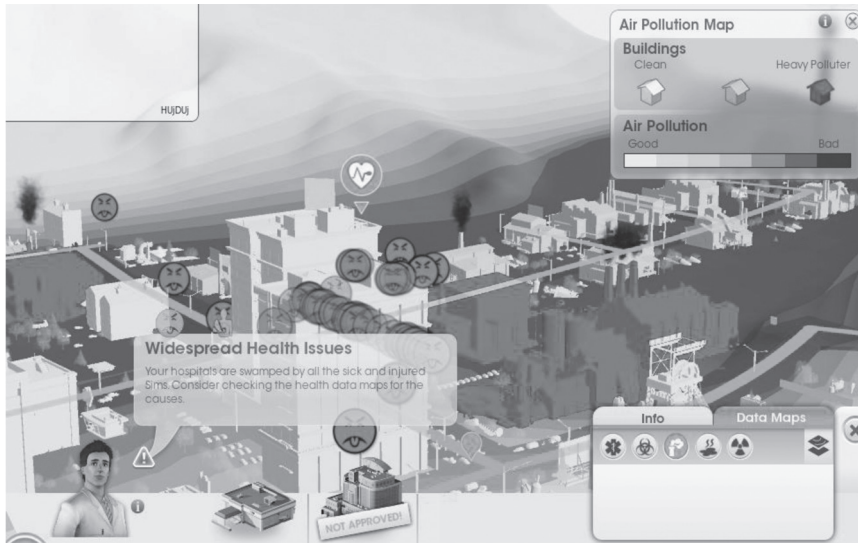
In the Jackson City challenge, the player (in the role of mayor) enters the city (Figure 1.1) and is told that residents seem unhappy and are leaving. Interaction with the Sim characters reveals that they are having trouble with air pollution. Players can explore data maps that show which buildings are 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

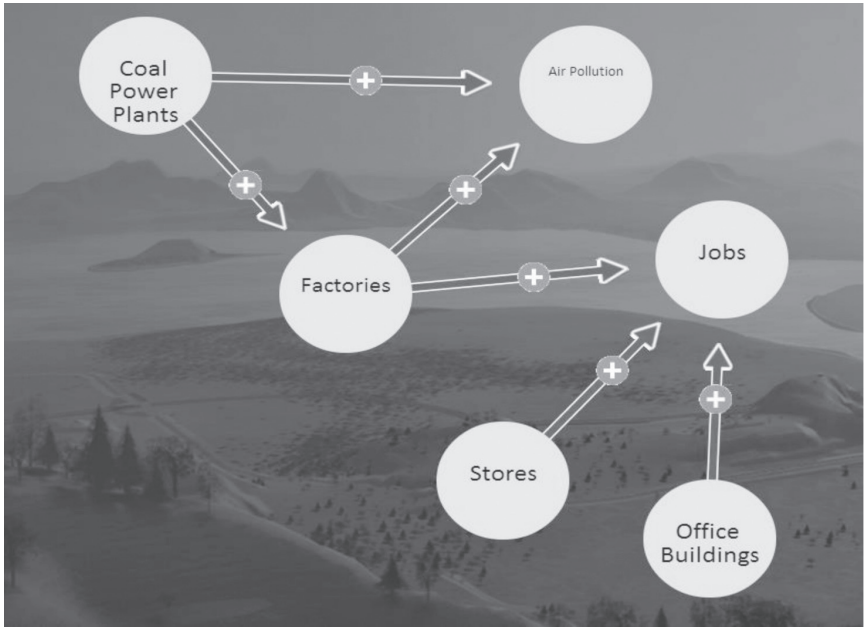


Source: Mislevy et al. (2014) (used with permission from the Institute of Play).

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

Obtaining, Evaluating, and Communicating Information

- Obtain and combine information from books and other reliable media to explain phenomena.

Disciplinary Core Ideas

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.

Crosscutting Concepts

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.

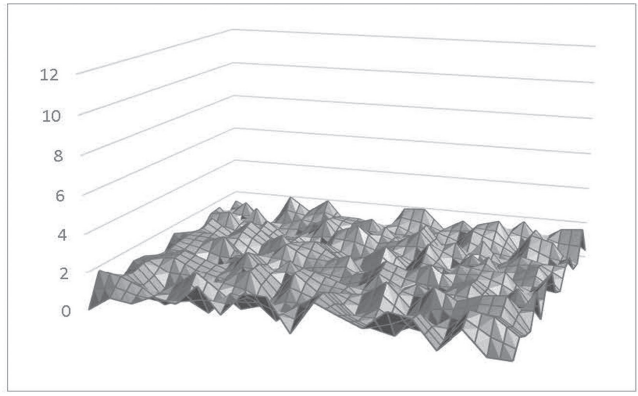
Panel (b) is the result after Carlos has worked through two in-depth investigations. We first notice spikes where resources have developed around the particular combinations addressed in these tasks with regard to core ideas and cross-cutting themes (as well as practices, representations, and so forth, 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).

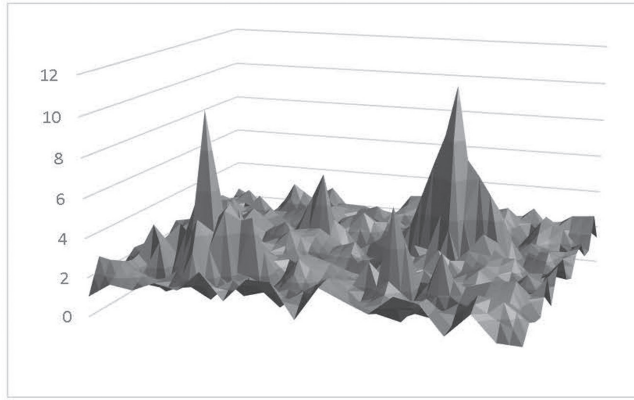
In addition to the peaks themselves, we also note ridges along the dimensions that were addressed. These represent resources that have developed in the experience that might have "hooks" that could be activated in some other contexts or with other disciplinary ideas. Carlos might encounter a new situation, 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.

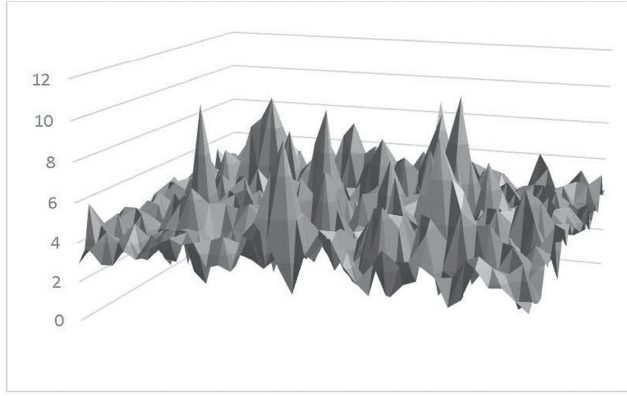
One point that will become important with regard to measurement is that Carlos's peaks and valleys are not in the same places as other students'. Some are similar; if Carlos and his classmates have all worked through SimCityEDU, they will have similarities in those regions where they have



(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.

5.3 Learning Progressions and Higher-Level Skills

Both learning progressions (Corcoran, Mosher, & Rogat, 2009) and higher-level/noncognitive/21st century skills (Darling-Hammond & Adamson, 2010) have been advocated for designing instruction and assessment. The sociocognitive perspective again offers insights into their nature as conceptual frames as distinct from individuals' capabilities, and sets the stage for subsequent sections' discussion of their role in different assessment use cases.

We can indeed identify concepts, methods, and strategies in activities that correspond to similarities across occurrences of what would, in everyday language, be called higher-order skills with a given name—problem-solving, creativity, collaboration, and so on. Whether a given person who is proficient at doing this in one context can do ostensibly similar things in another context is far from guaranteed. Research on learning from a sociocognitive perspective is helping us understand the reasons for sometimes-puzzling, seemingly conflicting, results from decades of study of transfer (Hammer et al., 2005). Although it is possible to identify cross-domain concepts, methods, and strategies at a more abstract level, we now better understand the results on higher-level skills that were becoming apparent decades ago. The conclusion Perkins and Salomon reached in their research synthesis in 1989 persists:

> Thinking at its most effective depends on specific, context bound skills and units of knowledge that have little application to other domains. To the extent that transfer does take place, it is highly specific and must be cued, primed, and guided; it seldom occurs spontaneously. The case for generalizable, context-independent skills and strategies that can be trained in one context and transferred to other domains has proven to be more a matter of wishful thinking than hard empirical evidence. (p. 19)

Developing higher-level resources that can be applied in new domains requires experience in particular contexts, generally several of them. It is facilitated by experience that makes the relationships between the abstractions and the particulars of contexts explicit to students.

This finding applies to learning progressions. A learning progression like the one in Table 1.1 is a good enough description of increasing levels of challenge in working with a given system, in a given context. This progression proved useful for designing the series of SimCityEDU challenges, each one posing a problem with additional complexities in the system at issue and requiring a certain kind of understanding to solve the problem. It was useful too for providing feedback to players within the game, and for structuring classroom discussions around the game experiences. It serves as an experience to help students develop higher-level resources that can

be applied beyond the context of learning. (See Songer, Kelcey, & Gotwals, 2009, on the value of learning progressions for higher-level skills more generally, in designing instruction and assessment as integrated with particular disciplinary content and activities.)

However, "variations among students' performances with respect to levels of learning progressions can show striking variability in different contexts and different content areas" (Sikorski & Hammer, 2010). Accordingly, in interpreting the Bayesian network measurement model in SimCityEDU, members of the design team urged a more local interpretation of the student-parameter values summarizing a student's play:

> This is not to say we believe a given student is "at" some particular level, even at a given point in time and with regard to a given system. A progression-level characterization of a person's systems thinking and actions can vary with the contexts and contents of situations. … [h]ow much performance varies, in what ways, and with what sensitivity to systems and contexts, is a central concern for inference about such broadly defined skills. (DiCerbo, Mislevy & Behrens, 2016, p. 257)

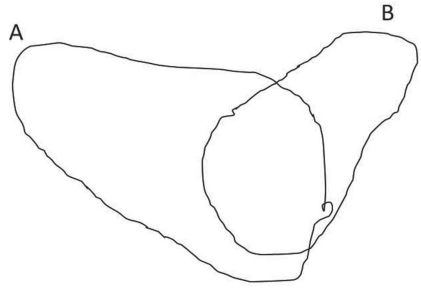
Section 10 follows up on this observation with reasons to develop conceptual frameworks and learning progressions for higher-level skills nevertheless. To anticipate, it agrees with the position taken by Songer, Kelcey, and Gotwals (2009) on their value in designing instruction and assessment as integrated with particular disciplinary content and activities—a strategy of which SimCityEDU is but one illustration.

6.0 IMPLICATIONS FOR LEARNING THAT HIGHLIGHT A SOCIAL ASPECT

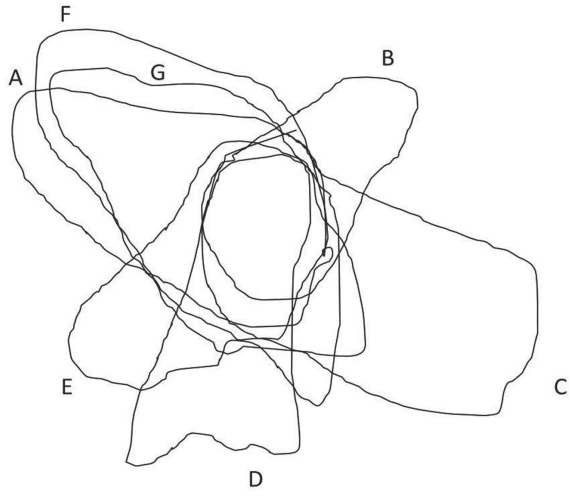
As mentioned, we learn from experience in unique situations structured around LCS patterns at many levels, and the resources we develop are initially tied to the circumstances of learning. The previous section looked more closely at cognitive aspects of this principle, with an eye toward school learning. This section looks more closely at social aspects, as to how the ways and extents that commonalities among peoples' trajectories of experience influence the resources they develop. Again, we will see in following sections how these ideas hold implications for assessment, and in particular for assessment with rich performance tasks. We will again use some simple sketches to suggest the ideas, and this time use vocabulary development as an illustrative context.

Panel (a) of Figure 1.5 represents in two dimensions the vastly richer milieus of experiences of two persons, Alicia and Bayar, growing up respectively on a farm and in a suburb in Illinois. Although they have never

interacted with one another, their experiences build around many common practices (going shopping, watching television, attending school, studying science in classes ostensibly under the same state-level standards, learning to speak and read in English, etc.), built around many common LCS patterns. This is represented by the overlap in their ovals. The resources they are developing are unique to them as individuals, but bear strong similarities as to many attunements, knowledge structures, and activity patterns. This includes for example much of the structure of language, seen not as a unitary, coherent, entity, but rather as recurring commonalities in peoples' interactions, varying from over time, over situations, and over people, but with enough commonalities to enable meaningful interaction.



(a) Alicia and Bayar.



(b) Alicia, Bayar,..., Federico, and Gina.

Figure 1.5. Hypothetical simplified milieu of linguistic, cultural, and substantive situations experienced by people.

The commonalities also include uses of many of the same words for similar purposes. Their understandings of words is unique, but quite similar for some, such as function words "is" and "because," due to the structural similarities of the many uses of these words as they have experienced them in contexts. It is more different for other words such as "cow," for which Alicia has a richer and more varied body of experience than Bayar, hence a richer, more varied network of resources to draw upon, whereas the reverse is true for "football." They have other experiences that are not at all similar, represented by the parts of their ovals that do not overlap. Alicia has developed resources associated with "skip loader" and Bayar has not, while Bayar has understandings of "nose tackle" that Alicia does not. Both have many and varied resources associated with the word "force," but because Bayar has taken an introductory physics course, his connects in some ways to the sense of the term as it is used in scientific communities.

Panel (b) adds more people, showing different amounts and locations of overlap and disparities. Note for example a greater overlap among Alicia, Federico, and Gina, all of whom grew up on farms and have begun studying agriculture at Western Illinois University. They are continually developing richer, more densely connected, more useful networks of associations with vocabulary connected with agriculture—not identical, partly due to their unique histories of experience, and partly because their motivations and diligence in their study differ.

Section 9 returns to these diagrams in connection with assessment design and assessment use cases. The general points to take from this discussion are these: Peoples' lexical resources develop through their experiences. Many elements are widely shared. More elements are more specialized to our communities, work, hobbies, and cultures. When there are strong similarities in experiences, people develop resources with corresponding similarities—many below the surface, to use the iceberg metaphor, that are simply presumed in instruction or assessment. They can impede learning or performance when a student's background experiences do not match up with the presumptions of an educational situation. The surrounding classroom support for SimCityEDU is meant to reduce some critical obstacles students would face. So are some initial tutorial challenges in the simulation environment, which help the player understand the interface and tools without yet pressing on the complexities of the pollution/jobs system.

7.0 ASSESSMENT DESIGN AND INTERPRETATION ARGUMENTS

This section reviews the structure of assessment arguments. It focuses on distinctions we will need for the evidentiary arguments that underlie

different assessment use cases. The same basic argument structure is used prospectively in designing assessment tasks and retrospectively in interpreting performance (Mislevy, 2006; Mislevy, Steinberg, & Almond, 2003). The structure adapts terminology and representations from Toulmin's (1958) general form, shown as Figure 1.6. Reasoning flows from *data* (D) to *claim* (C), justified by a *warrant* (W), which is supported by *backing* (B). The inference may need to be qualified *by alternative explanations* (A), which may be accompanied by *rebuttal* evidence (R) to support or weaken them.

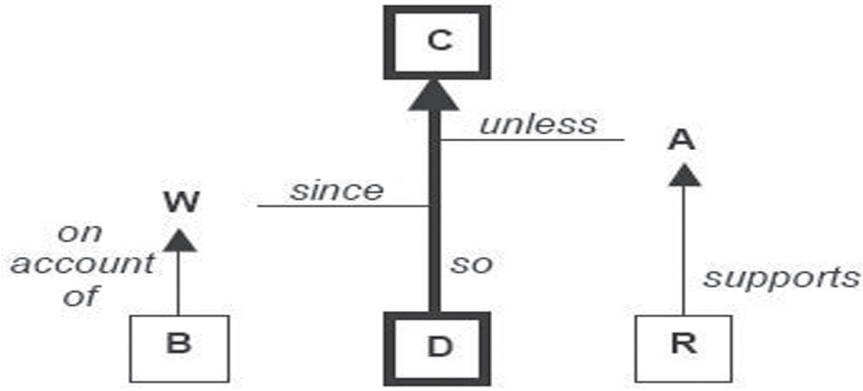


Figure 1.6. Toulmin's (1958) structure for arguments.

The elaborated interpretation argument structure is shown as Figure 1.7. At the top is the claim we want to make about a student, to be supported by evidence from her performance. A first important distinction among assessment use cases concerns the nature of the claim for a given use of an assessment. It might concern an individual student, for formative or for summative purposes, as opposed to being a nugget of evidence for a claim about the distribution of capabilities in a population. The claim may be conditional on certain information about the student or about the context of the intended score use. While the information associated with a claim is typically expressed in terms of a score of some kind, these conceptual aspects of a claim also determine its meaning. We will see that the same score, arising from the same performance, can take different meanings in different use cases.

At the bottom of Figure 1.7, shown as a cloud, is a student's performance in an assessment situation—a unique human action, from which we wish to identify what is meaningful for our purpose and map it into an across-student argument form. Supporting the claim in this manner are the first two types of data: features of the student's performance (often the only

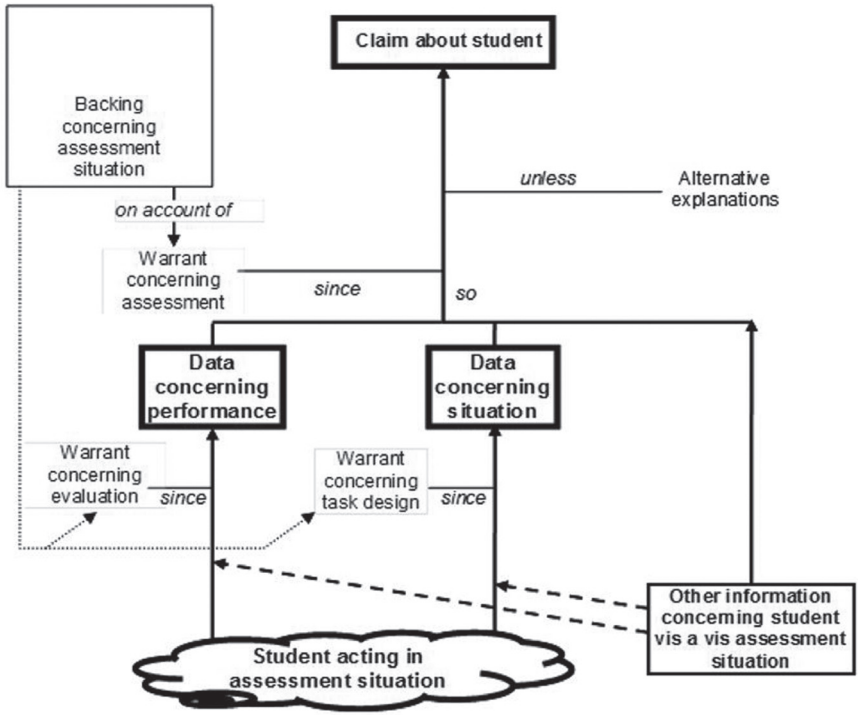


Figure 1.7. An assessment design/interpretation argument.

kind of evidence generally thought of as "the data" in assessment), but also features of the assessment situation. It is necessary to consider the two jointly to make sense of the performance, because warrants in assessment arguments address the kinds of things that students at different levels or with different configurations of proficiency are apt to do in what kinds of situations. Warrants in behaviorist assessment concern stimulus-response bonds and warrants in trait assessment concern tendencies toward behavior in broadly cast situations. Warrants cast in a sociocognitive perspective concern capabilities of context-dependent assembly of cognitive resources of many kinds in various situations. This is so even when the overt information is a simple score, the same one produced under the same psychometric model that could be used in a trait-based argument. A sociocognitive perspective, however, heightens a user's sensitivity to alternative explanations below the surface.

The third kind of data, "other information about the student vis a vis the assessment situation," is usually tacit in the visible machinery and procedures of an assessment, but it is equally critical to the interpretation of

a performance as evidence for a given claim. From the sociocognitive perspective, this is where some of the iceberg of the myriad resources below the surface that are needed for performance can become visible. It is a central concern in testing special populations, for example (Mislevy et al., 2013). In some uses, information about opportunity to learn or about students' cultural or linguistic background is important for interpreting their performances (Moss et al., 2008). It is critical in distinguishing the evidentiary value of performances in contextualized formative assessment uses and in assessments that "drop in from the sky," that is, assessments that have no predesigned connection to students' instructional, cultural, or personal backgrounds.

Alternative explanations in assessment are closely connected with validity (Messick, 1989). Of particular importance in performance assessment is the validity threat Messick calls "construct irrelevant variance": Poor performance can be caused by requirements for knowledge, activity patterns, or expectations that are needed to perform well but are not central to the intended interpretation. There are many of them, and often many are tacit. However, "other information" data may be available to remove looming alternative explanations. It may be information that is at hand, as by a teacher who has worked with a student for months. It may become known through the testing process, as when background information is gathered in student and teacher surveys in large-scale assessments. It may be created outside the assessment per se, as through practice problems or as when candidates must meet qualifications to take a test.

The critical elements in the argument for understanding the evidentiary characteristics of rich performance tests in different use cases are the nature of the claim, alternative explanations, and other information that may be available. None of them are visible in the assessment materials or in performances.

8.0 ASSESSMENT USE CASES

An educational assessment is used to gather information for some user(s), for some purpose, under some constraints. A user might be a teacher, a policymaker, or the students themselves. In some way, a user needs information about how educative efforts are faring in order to evaluate them, allocate resources, or plan next steps. The word "assessment" refers to a broad array of ways that actors gather information about students' capabilities—under different conditions, for different purposes, gathering data in different ways, and operating from different knowledge standpoints.

A "use case" in systems design describes the actors, information, and processes involved in meeting some recurring function, like withdrawing cash from an ATM or updating a customer database. A use case in assess-

ment describes a configuration of actors, information, and processes that serve a recurring assessment purpose in situations with recurring constellation of critical features. The definitions in this section and the discussion of their implications in the following section draws on Gorin and Mislevy (2013) and Mislevy and Duran (2014).

Table 1.3 addresses four (of many possible) use cases where one might use rich performance tasks, selected to highlight their paradoxical characteristics.

Table 3. Four Assessment Use Cases

Use Case	*Description*
1	**Formative assessment during learning activities** • Target claims: Finer-grained aspects of proficiency or performance, to support learning • Stakes: Low • Use of additional information: Very strong use • Contextualization of inference: Primarily to learning environment • Marginal vs. conditional: Highly conditional
2	**Summative assessment in a course of instruction** • Target claims: More coarsely-grained aspects of proficiency, to evaluate learning • Stakes: High • Use of additional information: Strong use • Contextualization of inference: Mostly to learning environment • Marginal vs. conditional: Predominantly conditional
3	**State-level accountability assessment** • Target claims: More coarsely-grained aspects of proficiency, to evaluate learning with respect to students, teachers, and/or educational systems • Stakes: High for at least some level • Use of additional information: Low use • Contextualization of inference: To learning environment • Marginal vs. conditional: Predominantly marginal
4	**Large-scale educational survey (e.g., NAEP)** • Target claims: More coarsely-grained aspects of proficiency, to provide feedback on educational systems at the level of populations and study relationships between these proficiencies and covariates • Stakes: Low • Use of additional information: Moderate use • Contextualization of inference: To learning environment • Marginal vs. conditional: Both marginal and somewhat-conditional inferences

- Use Case 1: Formative assessment during learning activities
- Use Case 2: Summative assessment in a course of instruction
- Use Case 3: State-level accountability assessment
- Use Case 4: Large-scale educational survey

These use cases differ from one another with respect to one or more dimensions that affect the evidentiary value of data from performance assessment. The following terms appear in the table. The next section provides more discussion and examples.

The *Target claims* indicate who is being assessed, at what grainsize the inference is being made, and what the main purpose(s) of the assessment are. However, a fuller understanding of the nature of the claim in the argument involves the categories listed below as contextualization of inference, use of additional information, and the degree of marginality versus conditionality of the inference.

The *Stakes* of an assessment concerns the consequences of the results. Low stakes mean low consequences, as is often the case in formative assessment (at least outside the classroom; Shepard (2008), has pointed out that consistent errors in formative assessment in the classroom can seriously erode students' opportunities to learn). High stakes uses can affect grades, graduation, or licensure for individuals, evaluation for teachers, or funding for educational systems. An assessment can be low stakes at one level but high at another, such as a statewide test that affects school governance but has no specific consequences for individual students.

Use of additional information concerns the degree to which local information about the relationship between assessment tasks and students' backgrounds is used in inference, with respect to instruction, culture, language, disability, etc.

Contextualization of inference concerns the degree to which a claim extrapolates to situations beyond the immediate assessment situation. In generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) it corresponds to the definition of the universe score in the interpretation argument. This issue is particularly salient in performance assessment (Messick, 1994).

Conditional versus marginal inference concerns the degree to which claims are made conditional on the additional-information characteristics of students.

We now turn to the details of the use cases, seeing how these characteristics affect the evidentiary value that rich performance tasks afford.

9.0 EVIDENTIARY CHARACTERISTICS OF PERFORMANCE TASKS

The preceding sections have developed the concepts we need to discuss the evidentiary characteristics of rich performance tasks in different assessment use cases. This section walks through the four use cases described in Table 1.3.

Use Case 1: Formative Assessment During Learning Activities

This use case is the poster child for rich performance tasks. It can capitalize on all of the advantages that advocates claim, and avoid the disadvantages others caution against. When used as an integrated component of learning, performance tasks can be selected, constructed, or sequenced to match learners' backgrounds with respect to many of the necessary but ancillary aspects of the situation. This intentional congruence provides for rich and concrete instantiation of higher-level skills without becoming overwhelming (recall Section 5.3). The users of the information can be teachers, the learners themselves, or both—informing larger feedback loops for a teacher, to help guide classroom discussions or feedback to individuals, and tighter loops for students closer to their actions. Especially for more extended tasks, it is important that the activity and the value of the formative feedback be sufficiently contextualized to students' learning to justify the time that is spent on this rather than on other activities. Stakes are low, because the feedback cycles are tight, quick, and frequent, and consequences of errors are small and easy to recover from (as long as they are not cumulative, and as long as the teacher or student knows what to do about them, as Shepard, 2008, points out). Several evidentiary implications follow from this deep contextualization.

The claims in the assessment argument address understanding and action in the rich situation at hand, focusing on knowledge, practices, or themes that are the target of learning. Because each assessment situation is integrated into a larger learning situation, a great deal can be known about the student's background with respect to many of the nonfocal LCS patterns that are involved. The assessor's knowledge of these matches attenuates many alternative explanations that would otherwise weaken inference.

Figure 1.8 suggests this effect using the overlapping ovals from Figure 1.5. The double-circled shapes represent three students who are all working through SimCityEDU in their classroom, receiving support and practice that lets them focus on the systems aspect. The dashed oval represents the shared experience within which the formative assessment takes place. The stars represent assessment occasions in, say, the Jackson City challenge. Ruling out many alternative explanations through both the support and

the users' knowledge about the support increases the validity of inferences. In particular, test fairness increases when local adaptation reduces demands that are associated with cultural factors, linguistic backgrounds, and disability status (Mislevy & Duran, 2014; Mislevy et al., 2013). In the context of language testing, Swain (1985) used the term "biasing for the best" for adapting nontargeted aspects of performance tasks to examinees.

Figure 1.8. Assessment occasions in formative assessment in a performance task meant to develop learning in a given context.

The primary claims in learning tasks concern students' understanding and action in the situation at hand; extrapolation is not intended. Recall the discussion of formative assessment for Carlos in Section 5.1. SimCityEDU; formative assessment of his systems-thinking at that moment is *conditional* on his experience so far with the SimCity environment and the Jackson City jobs-and-pollution system in the challenges. This is suggested in Figure 1.9, where the circled peak represents the content and context of the learning task as well as systems-thinking resources. In generalizability theory terms, both the task space and the universe of generalization are focused quite narrowly. Many facets of a conceivable universe of rich performance task are fixed: the simulation environment, the system at issue, the representations, even the level of systems thinking that is required in the challenge. What is more, they are fixed at values where it is known, by the way the task is designed and exactly when it is used, that the student has already developed

many relevant resources that can be brought to bear in the task that pushes on only a few facets that are the current target of learning.

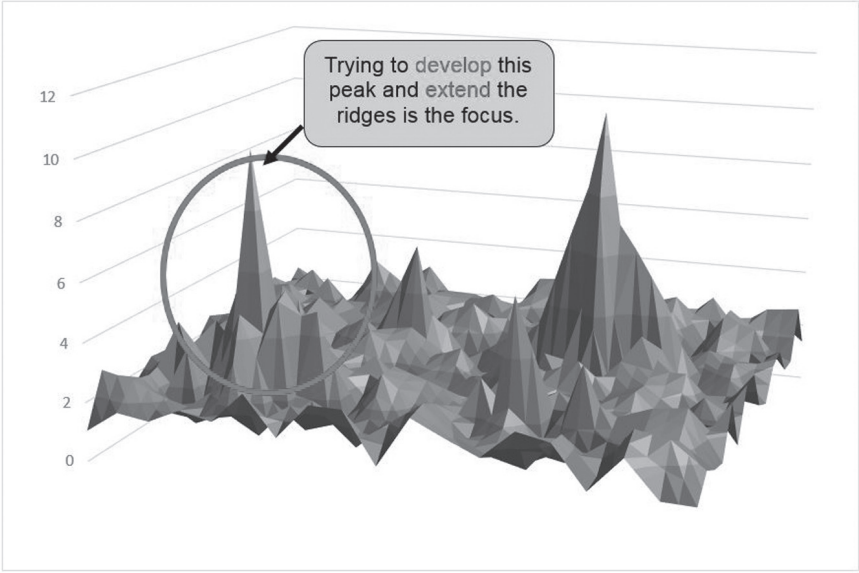


Figure 1.9. The target of claims in formative assessment in a performance task meant to develop learning in a given context.

A corollary of these measurement circumstances is that the evidentiary value of such task is not simply inherent in the task itself. Rather, it emerges in its specifically targeted use for students known to have had some particular kinds of experiences and learning, at a very particular time, for very particular assessment purposes. Assessment and feedback here is conditional on the context and content. It is phrased, however, using some of the more abstract language and concepts of systems thinking. This framing helps Carlos and the other students at this point develop resources that will be useful beyond this game and these situations. Succeeding in this effort corresponds to increasing the height of the surface along the ridge representing systems-thinking resources.

Although the experience is designed to foster the development of more general systems-thinking schemas and resources, it is not of immediate concern whether a student's systems thinking in SimCityEDU would predict their understanding and action in a wolves-and-moose scenario, or some other context and system sampled from a system-thinking task domain. That would correspond to a broader universe of generalization, and could be described as *marginal* inference from the same data. These latter kinds of claims are of interest nevertheless, because we really do want students to

develop resources that will be useful beyond the SimCity world. They will be addressed in the next two use cases.

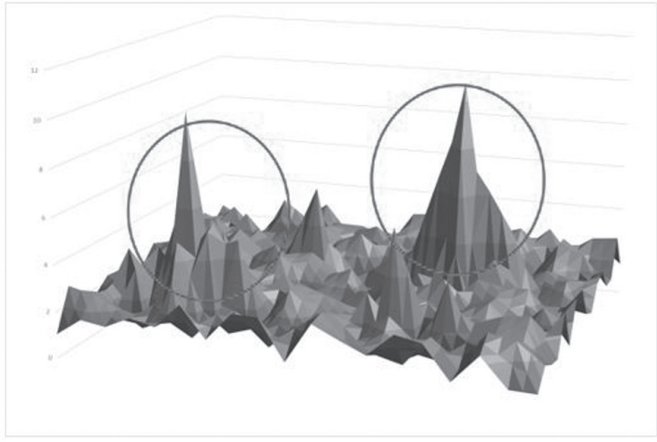
Use Case 2: Summative Assessment in a Course of Instruction

This use case concerns assessment that is again integrated with a course of learning, but now used summatively. Stakes are higher than in formative assessment: a course grade or a certificate, for example, or the opportunity to move to a subsequent course. Suppose students have worked through both SimCityEDU and another simulation-based systems-thinking unit, based on say NGSS performance expectation MS-LS1-7, "Develop a model to describe how food is rearranged through chemical reactions forming new molecules that support growth and/or release energy as this matter moves through an organism." As in the previous use case, the assessor has a great deal of information about the students' background experiences on which to define claims, design tasks, and rule in or rule out alternative explanations.

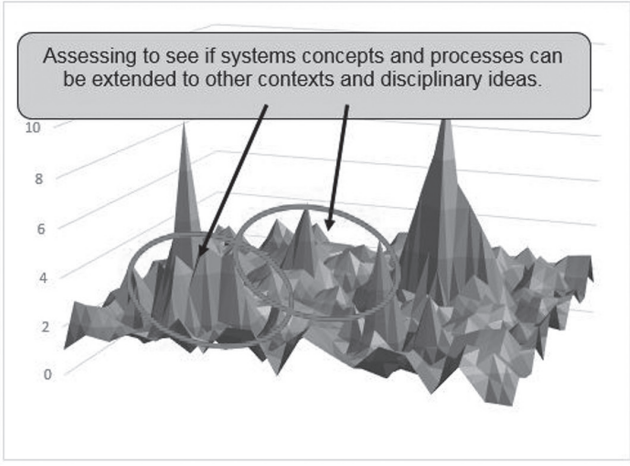
Performance tasks of two kinds might be devised for an end-of-course assessment. First, students can be presented new challenges within the now-familiar SimCityEDU pollution world or the equally familiar molecules-in-cells world. The claims addressed by such tasks concern a student's capabilities within contexts and using disciplinary ideas that *the assessor knows the students are familiar with*. In particular, they probe the students' capabilities in applying the systems concepts and tools within these familiar contexts. This "close transfer" of universe of generalization is shown as Figure 10(a). A "far transfer" universe of generalization for a claim about the extent to which a students could apply systems-thinking might utilize a context and disciplinary ideas other than the ones used in instruction, as suggested in Figure 10(b). These tasks would use contexts and disciplinary ideas *the assessor knows the students are* not *familiar with*. Other aspects of the transfer tasks, perhaps such as systems interfaces, response expectations, and diagramming tools could be made similar, so difficulties on these counts would not be viable alternative explanations.

To the outside observer, there is no distinction between the familiar and the unfamiliar tasks. Their different evidentiary value for claims about near and far transfer only exists because the assessor can incorporate information about the different relationships between the students and the tasks into the argument.

Because this use case involves higher stakes for students, accuracy of inferences matters more. Whereas providing formative feedback during performance on a learning task addresses claims local to that task,



(a) Neighborhoods closely related to the context of learning



(b) Neighborhoods involving systems thinking but with contexts and disciplinary ideas farther from the context of learning

Figure 1.10. Targets for claims concerning systems thinking in summative assessment.

course-level summative claims concern performance over classes of tasks (perhaps near-transfer tasks, medium-transfer tasks, and far-transfer tasks, all defined as relative to the course of instruction). Even though much is known about students for avoiding alternative explanations and even though tasks are integrated with their course of learning, research since the 1980s shows repeatedly that person-by-task variability is a large component

of variation in performance tasks (Linn, 1994). Studies that examine it even show that person-by-task-by-occasion variance can be surprisingly high too (Ruiz-Primo & Shavelson, 1996). The same student working on the same task a few weeks later can perform quite differently.

In generalizability theory, the generalizability coefficient extends the familiar reliability coefficient by taking into account the effects of multiple sources of uncertainty and of the number and configurations of tasks and raters. When the target of inference is an average score over some domain, we are in effect trying to estimate the average height of a student's topographic map. The variance components mentioned above characterize the irregularity of the surface. They appear in the denominator of the generalizability coefficient. The more extreme the peaks and valleys are, the less information any one score provides. One must average over a larger number of tasks as needed to obtain a given accuracy. A large number of tasks is therefore usually needed to obtain reliable scores (Shavelson, Baxter, & Gao, 1993). For example, the National Board of Medical Examiners uses 13 tasks in computer-based patient-management examination to achieve sufficiently reliable scores in physician licensure. This test takes a full day. Together with a day-long multiple-choice test, the current fee is close to $1,000.

In light of this result, one useful design strategy is to have more but shorter performance tasks in assessments used for purposes such as final grades or certifications. While the Cisco Networking Academy uses simulation-based troubleshooting tasks that might take an hour to work through during learning (Use Case 1), a course final exam or a licensure test contains smaller, more focused, slices of several such tasks.

Use Case 3: State-Level Accountability Assessment

Use Case 3 represents the uses of rich performance tasks that most strongly merit the cautions Linn (1994) summarized. The cautions are well deserved, even if the tasks are the same as the ones used in the felicitous Use Case 1.

Consider a state-level accountability test, used for high-school graduation at the level of students. Suppose the state has adopted the NGSS, so systems thinking, the inquiry practices, and disciplinary ideas in Jackson City are all within the expectations held for the students in the state. So too are other, similar performance tasks like the wolves-and-moose and the food-energy-system tasks. Any could appear on the assessment of any student in the state. Any student in the state may or may not have had in-depth experience with the kinds of interfaces, contexts, or particular systems at issue.

A key difference from the previous use case is that now *the assessor does not know how any students' background experiences and task demands match up*. The stars in Figure 1.11 represent a sample of tasks across a broad domain defined by contexts, disciplinary ideas, practices, and cross-cutting themes, shown on a typical student's topography of capabilities.

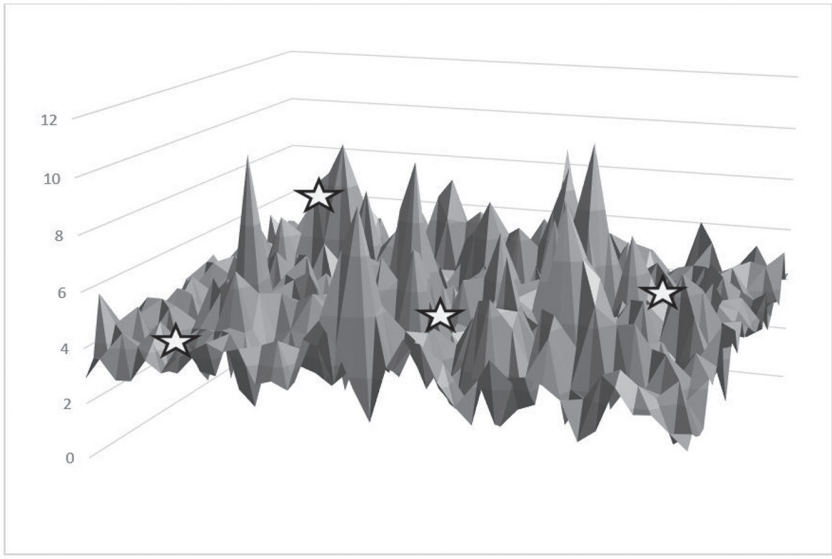


Figure 1.11. A sample of performance assessment tasks in a complex domain, matched against one student's topography of capabilities.

It may be of interest to know how well a student can, say, carry out systems thinking across such a broad universe of generalization. As in the previous case, the target "universe score" is the average height of an irregular surface. But now the surface is larger, and the population is more diverse with respect to the mixes of LCS patterns in each of the tasks. There are more potential alternative explanations for poor performance, because unlike the classroom teacher, the user does not possess the additional information to rule out as many of them. Because performance tasks usually take more time, few can be administered. The contribution of the person-by-task variance component is larger and the generalizability coefficient is lower.

This then is the central paradox. When contextualized with instruction (Use Case 1), richer and more integrated tasks provide better conditional information to advance individuals' learning; but when they are not contextualized (Use Case 3), the greater person-by-task interaction variance degrades marginal inference about individuals.

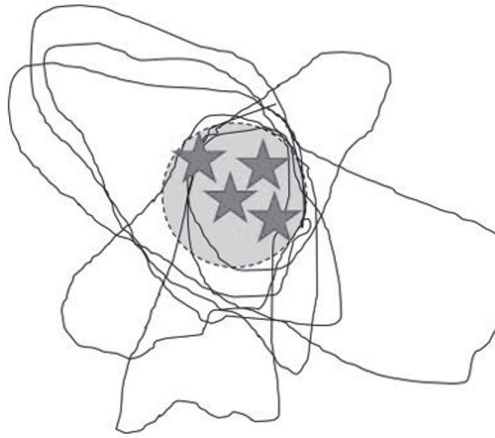
Further, we recall from Section 5.1 that tasks are more difficult in a marginal sense when they employ more challenging use of practice, more advanced disciplinary knowledge, more cross-cutting ideas involved in more subtle ways (even though the use of such tasks is well-targeted conditionally when they are integrated in learning experiences, as discussed above in connection with Use Case 1). There are more ways a student can experience difficulties; more facets of tasks are free to vary. These facets are not fixed at targeted levels as they were in Use Case 1, with tasks selected at the time of use expressly so these facets would not be significant sources of difficulty.

What's more, the greater the diversity of the backgrounds of students is, the stronger the effect on person-by-task variance component with tasks selected without targeting will tend to be, and the lower the generalizability coefficient for marginal inferences for individuals will be. Alas, these are the targeted claims in Use Case 3.

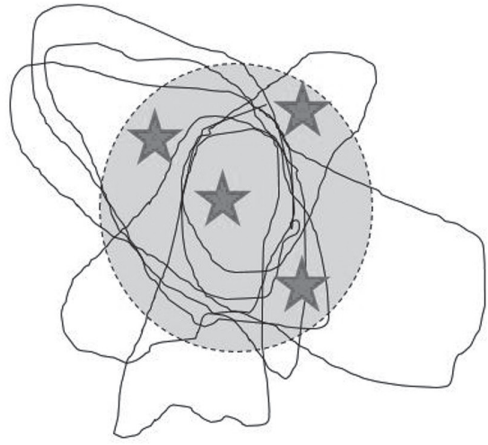
The situation is illustrated in Figure 1.12. Panel (a) shows a test comprised of performance tasks, but ones which press less hard on practices, disciplinary knowledge, or cross-cutting themes. Their contents are more likely to have been encountered by more of the testing population. The harder a designer presses along any of these dimensions, the more outside previous experience a task will be in some aspect(s) for some examinees, but, as Panel (b) suggests, different ones. There is thus a larger person-by-task interaction variance component when the inference concerns a students' systems-thinking across a broad domain of performance expectations tasks.

Use Case 4: Large-Scale Educational Survey

This use case is exemplified by large-scale educational surveys such as the National Assessment of Educational Progress (NAEP) in the United States, the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). Samples of students within or across countries are administered assessments that usually include performance tasks, for the purposes of surveying achievement in these jurisdictions and supporting research on its correlates. It is similar to the previous use case in that tasks are administered to students about whom relatively little is known, usually just from background surveys of the examinees or school officials. It differs as to the intended claims: Not inferences about individuals, but about distributions of performance in jurisdictions and subpopulations, about correlates of this performance with background variables, and in some cases, in-depth looks at patterns of performance over examinees in particular tasks.



(a) Sampling simpler tasks with elements more likely to be shared.



(b) Sampling richer tasks with elements less likely to be shared.

Figure 1.12. Sampling domains of less ambitious and more ambitious performance tasks.

When it comes to addressing higher-level skills such as systems thinking in this use case, the same pictures for Use Case 3 apply, namely Figure 1.11 and Figure 1.12(b). There is limited, and often no, adaptation or selection of tasks to students' backgrounds in order to optimize inference about individual students. Indeed, administering tasks without adaptation provides better evidence about the distribution of performance that would have been observed had the same tasks been administered to everyone in

the population—even for tasks that would have been meaningless to many of them! Just learning this is one of the kinds of things such as survey is meant to reveal.

Under these conditions, student-level generalizability coefficients would be abysmal for measuring a higher-level skill such as systems thinking as defined by a broad domain of performance tasks that involve interactivity, rich contexts, ambitious disciplinary content, and cross-cutting themes. Each individual might take one or two tasks sampled from this broad domain, randomly matched (or mismatched) with her jagged topograph. But satisfactory accuracy can be obtained nevertheless for inferences about a group's distribution for a measure of a variable so defined, as long as the samples of students and tasks are large enough (Pandey & Carlson, 1976).

Furthermore, if enough students are administered a given task and sufficiently rich data are obtained (e.g., a detailed log file of their actions), the data provide strong evidence about claims concerning how samples of students perform within each particular task: their choices, the way they use tools, the steps they take, where they run astray, where they run into problems, and how they respond in places in the task where specific responses are required (e.g., filling in a representation). This information can be correlated with whatever background information may be available. Data from rich performance tasks in Use Case 4 provide good evidence for these kinds of claims, even though they provide poor evidence about individuals for broadly-cast skills. We may also discover patterns across tasks—not about individual students, but about ways people think through and interact with such tasks.

10.0 THOUGHTS ON USING PERFORMANCE TASKS

Rich performance tasks and multiple-choice test items are but two of many ways to gather information about students' capabilities (Scalise & Gifford, 2006). It is the job of assessment developers to understand the design space—all the evidentiary, logistic, and educative characteristics of different assessment types—and propose assessment configurations that suit given contexts and purposes. Rich performance assessments are currently of interest partly because of developments in understanding how people learn, but even more because of advances in technology. Game- and simulation-based assessments, for example, enable us to observe, evaluate, and provide feedback on interactive performances in complex environments that until recently could have only been done at small scale, at great costs, or with questionable reliability. The production possibility frontier for educational assessment has been pushed outward.

Yet no matter how sophisticated, integrated, and automated a task might be, it may still not provide good evidence for some purposes, in some situations, with some states of users' knowledge. It is the characteristics of assessment tasks in conjunction with these contextual factors that determine the evidentiary value of rich performance tasks, or indeed any others. This short section offers some observations on ways that researchers and practitioners are finding to optimize the use of rich performance tasks in assessment. It does not focus on technology, even though it is advancing rapidly and opening new possibilities to every facet of design, delivery, and use of data (for a few examples, see Behrens & DiCerbo, 2014; Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012; Luecht, 2013; and Sottilare, Graesser, Hu, & Brawner, 2015). The focus here is evidentiary issues, with particular attention to the expanse of LCS patterns involved in rich performance tasks and the match between tasks and students in this regard.

Practices such as inquiry, higher-level skills such as systems-thinking, and cross-cutting themes such as energy transfer are resemblances across ways of thinking and acting in many possible contexts with different disciplinary and social particulars. Learning progressions and assessment design patterns are two tools that bring out these regularities to support instructional design and assessment development across particulars. This is so even though they don't work as well to define "traits" across wide domains of tasks in diverse populations (Songer, Kelcey, & Gotwals, 2009).

Assessment design patterns (Liu & Haertel, 2011; Mislevy, Riconscente, & Rutstein, 2009) describe at a higher level of abstraction the elements of tasks for assessing a higher-level skill such as systems thinking or model-based reasoning, as they can be fleshed out with modes of assessment, for different purposes, and with particular content. They are organized around assessment arguments, and highlight the roles of the other demands that will be present in a task, and ways of matching up with or sampling across students' backgrounds in these regards. Design patterns not only help assessments developers create tasks for large-scale assessments, but they help teachers adapt task schemas to local information about their students, and they support principled adaptation of tasks to diverse populations (Haertel, DeBarger, Villabla, Hamel, & Colker, 2010).

Learning progressions (Alonzo & Gotwals, 2011; Corcoran, Mosher, & Rogat, 2009) like the one in Table 1.1 are useful in a similar way, as was discussed in connection with the design of SimCityEDU. They can additionally be used in creating or selecting tasks that help match students' backgrounds for integrated tasks.

Consider a design space of tasks that integrate systems thinking, plant respiration, and reading in English, and suppose there is a learning progression available for each of these broad dimensions. If we (say as his

teacher) roughly know a student Daquan's typical level of performance on each dimension in turn, *conditional on capabilities with whatever other demands happen to be in a situation*, we can better aim a task for him from this space. Suppose his skill levels in this particular way of defining them are L_{ST}, L_{PR}, and L_{RinEng}. We could create a task that probes at systems thinking, for example, by casting the plant respiration substance at level L_{PR}-1 and the reading demand at level L_{RinEng}-1, but the complexity of the system at level L_{ST} or L_{ST}+1 (Mislevy & Duran, 2014). We know of course that operating at given levels on multiple progressions when other aspects of the situations are familiar does not ensure a student will be able to perform comparably in a situation that poses new combinations of elements at those levels. Similarly, a combination above a student's typical levels of performance with respect to multiple progressions may happen to coincide with an island of expertise he happened to have developed. But this strategy does improve the odds by avoiding combinations we can expect a forehand to be problematic.

The learning progressions and design patterns strategies require a principled approach to understanding the abstracted patterns in disciplinary knowledge, practices, and cross-cutting themes. The idea applies across disciplines. These are powerful organizational structures to help us design of instruction and assessment. But for the reasons discussed above, they need not correspond with organizational structures in students' minds, and they need not lead to satisfactory overarching constructs to assess. Over time, with practice, through many situated and contextualized experiences, experts do develop sets of resources that are reflect the knowledge structures and activities in a domain mapping (Glaser & Chi, 1988). Even then, large person-by-task variation exists as tasks push out to ever more specialized subdomains and unique combinations of contexts and practices.

Understanding the strengths and weaknesses of different assessment approaches suggests strategies that combine approaches in a more encompassing system. High-stakes usage of NBME's simulation-based cases in the medical licensure sequence, for example, appears only after a medical student has passed the multiple-choice examinations earlier in the sequence and experienced simulation-based cases in medical school and practice sessions. Many alternative explanations for poor performance will have been weakened by this point, and much can be presumed about their resources for at least some aspects of medical knowledge and skills. Even so, considerable person-by-task variation remains in the simulation-based case assessment.

In the 1990s, the California Learning Assessment System (CLAS) envisioned a combination of very different kinds of assessment tuned to different aspects of learning (Knudson, Hannan, & O'Day, 2012). Portfolios of local work would provide data from Use Case 1 with few

constraints. Curriculum-embedded assessment would come with learning support; this was Use Case 1 also, but with more common context supplied in order to remove some alternative explanations and provide stronger cross-locality comparisons. On-demand assessment was an instance of Use Case 3. CLAS ended before it was fully implemented, due in part to social and political factors but also due to low generalizability problems with student-level writing scores from the on-demand portion (Cronbach, Bradburn, & Horvitz, 1994)—exactly the issue discussed above under Use Case 3.

Some additional approaches to combining contextualized assessment locally for learning (Use Case 1) and broader evaluation for moderate stakes (Use Case 2) are discussed in Mislevy (2008). The Studio Art portfolio assessment from the Advanced Placement program is discussed as an example of one of a number of configurations that could use data for different inferences at different levels of an assessment system.

11.0 CONCLUSION

Research from the learning sciences reinforces the value of rich performance tasks in students' learning. What is their value as tools to assess learning? The answer cannot be determined from the form and the processes of tasks alone. Sorting through the value of the information for a given purpose requires an accounting of what user needs information, for what purpose, how the tasks relate to the examinees' backgrounds, and what the user knows about the relationship. Concepts from sociocognitive psychology, evidentiary reasoning, and psychometrics (particularly generalizability theory and generalizations thereof) play useful roles in sorting through the details in a given application. The general results discussed in this chapter are consistent with the history of research in performance assessment:

- Rich, complex, performance tasks are well-suited to learning and to assessing individuals when contextualized with respect to targets of learning and students' experiential backgrounds. This can be done for formative purposes (Use Case 1) and summative purposes (Use Case 2). These cases are similar in their contextualization, but with the latter having a desire for broader inferences and a need for greater accuracy, thus requiring more tasks.
- They are not well-suited to assessing individuals when the inference is not contextualized with respect to targets of learning and students' backgrounds, increasingly so as the examinee population is more diverse. (Use Case 3)

- They are well-suited to surveying populations and studying the relationships of performances in particular tasks with regard to targets of learning and students' experiential backgrounds. (Use Case 4)

The very same contextualization that strengthens inference within contexts and contents also contributes construct-irrelevant variance for inferences that extend to other contexts and contents. Advances in technology can make any configuration of assessment richer, cheaper, more interactive, easier to evaluate, and stronger in the information it provides. But these improvements take place within the basic evidentiary-reasoning structure of a given situation, which define the possibilities and the limitations for inferences that can be drawn.

ACKNOWLEDGMENTS

This chapter is based on a presentation at the Fifteenth Annual Maryland Conference: Test Fairness in the New Generation of Large-scale Assessment, October 29–30, 2015, at the University of Maryland, College Park, MD. I am grateful to Maria Elena Oliveri and Leslie Nabors Olah for many helpful comments on an earlier version.

NOTES

1. Why Debate in Class? Downloaded February 23, 2016, from http://www.sas.upenn.edu/cwic/docs/db1.doc
2. Low generalizability means that a student's score on one task, or as evaluated by one rater, or as performed at one occasion, does not convey very much information about how the she would score under a different, equally acceptable, configuration that might have been used. Generalizability coefficients are generalizations of reliability coefficients that can encompass multiple facets of behavioral observations, such as students, tasks, raters, occasions, formats, and so on, and the variability associated with each (Cronbach et al., 1972). When a generalizability is low, one needs more tasks, raters, or more of whatever sources of variation are reducing generalizability, to obtain a given level of accuracy.
3. Applied, to be sure, but sometimes in ways that result in confusion or hinder learning. Hill and Larsen (2000), for example, carry out in-depth conversations with children about their thinking as they responded to reading comprehension test items, and found how subtle

differences in the cultures and language patterns in children's homes and neighborhood often added meanings that were quite logical, yet sometimes counter to the meanings the developers had intended.

4. Attributed to Gee by Dan Hickey at http://remediatingassessment.blogspot.com/2010/01/can-we-really-measure-21st-century.html

REFERENCES

Alonzo, A. C., & Gotwals, A. W. (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam, The Netherlands: Sense.

Behrens, J. T., & DiCerbo, K. E. (2014). Technological implications for assessment ecosystems: Opportunities for digital technology to advance assessment. *Teachers College Record, 116,* 1–22.

Branham, R. J. (2013). *Debate and critical analysis: The harmony of conflict.* New York, NY: Routledge.

Bransford, J. D. & Schwartz, D. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education*: *Vol. 24* (pp. 61-100). Washington, DC: American Educational Research Association.

Brown, N.J.S. (2005). *The multidimensional measure of conceptual complexity* (Tech. Rep. No. 2005-04-01). Berkeley, California: University of California, BEAR Center.

Cheng, B. H., Ructtinger, L., Fujii, R., & Mislevy, R. (2010). *Assessing systems thinking and complexity in science* (Large-Scale Assessment Technical Report 7). Menlo Park, CA: SRI International. Retrieved from http://ecd.sri.com/downloads/ECD_TR7_Systems_Thinking_FL.pdf

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education*, *42*(3), 309–328.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (Research Report No. RR-63). New York, NY: Consortium for Policy Research in Education.

Cronbach, L. J., Bradburn, N. M., & Horvitz, D. G. (1994, July). *Sampling and statistical procedures used in the California Learning Assessment System* (Report of the Select Committee). Sacramento, CA: California State Department of Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Crowley, K., & Jacobs, M. (2002). Islands of expertise and the development of family scientific literacy. In G. Leinhardt, K. Crowley, & K. Knutson (Eds.), *Learning conversations in museums* (pp. 333–356). Mahwah, NJ: Lawrence Erlbaum.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.* Palo Alto, CA: Stanford Center for Opportunity Policy in Education (SCOPE), Stanford University, School of Education.

Davey, T., Ferrara, S., Holland, P. W., Shavelson, R. Webb, N. M., & Wise, L. L. (2015). *Psychometric considerations for the next generation of performance assessment.* Princeton, NJ: K–12 Center at ETS.

DiCerbo, K., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R.J., Bauer, M., & Jackson, T. (2015). The role of exploratory data analysis in the development of game-based assessments. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 319–342). New York, NY: Springer.

DiCerbo, K., Mislevy, R. J., & Behrens, J. T. (2016). Inference in game-based assessment. In H. F. O'Neil, E. L. Baker, & R. Perez (Eds.), *Using games and simulations for teaching and assessment* (pp. 253–279). New York, NY: Routledge.

Dillon, G. F., & Clauser, B. E. (2009). Computer-delivered patient simulations in the United States Medical Licensing Examination (USMLE). *Simulation in Healthcare, 4,* 30–34.

Fauconnier, G. (1999). Methods and generalizations. In T. Jannsen & G. Redeker (Eds.), *Cognitive linguistics: Foundations, scope, and methodology* (pp. 95-127). Berlin, Germany: Mouton de Gruyter.

Fauconnier, G., & Turner, M. (2002). *The way we think.* New York, NY: Basic Books.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Hillsdale, NJ: Erlbaum.

Gobert, J. D., Sao Pedro, M., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 5,* 153–185.

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment.* Princeton, NJ: K–12 Center at ETS.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53,* 5–26.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–47). New York, NY: Simon & Schuster Macmillan.

Haertel, G., DeBarger, A. H., Villabla, S., Hamel, L., & Colker, A. M. (2010). *Integration of evidence-centered design and universal design principles using PADI, an online assessment design system* (Technical Report 3). Menlo Park, CA: SRI International.

Haggard, P. (2005). Conscious intention and motor cognition. *Trends in cognitive sciences, 9,* 290–295.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 89-120). Greenwich, CT: Information Age.

Hill, C., & Larsen, E. (2000). *Children and reading tests. Advances in discourse processes* (Vol. 65). Westport, CT: Greenwood Publishing Group.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York: Cambridge University Press.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92,* 109–129.

Knudson, J., Hannan, S., & O'Day (2012). *Learning from the past: Drawing on California's CLAS experience to inform assessment of the Common core*. Washington, DC: American Institutes for Research. Retrieved from January 31, 2016, from http://www.cacollaborative.org/sites/default/files/CA_Collaborative_CLAS.pdf

Lee, C. D. (2008). Cultural modeling as an opportunity to learn: Making problem solving explicit in culturally robust classrooms and implications for assessment. In P. A. Moss, D. Pullin, E. H. Haertel, J. P. Gee, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 136–169). New York, NY: Cambridge University Press.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, DC: American Council of Education.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

Liu, M., & Haertel, G. (2011). *Design patterns: A tool to support assessment task authoring* (Large-Scale Assessment Technical Report 11). Menlo Park, CA: SRI International.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology, 14*. Retrieved January 31, 2016 from http://www.jattjournal.com/index.php/atp/article/view/45254

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Phoenix, AZ: Greenwood.

Mislevy, R. J. (2008). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. In P. A. Moss, D. Pullin, E. H. Haertel, J. P. Gee, & L. J. Young (Eds.). *Assessment, equity, and opportunity to learn* (pp. 259–294). New York, NY: Cambridge University Press.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 59–81). New York, NY: Springer.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., John, M., Bauer, M. I., Hoffman, E., von Davier, A. A., Hao, J. (2014). *Psychometric considerations in game-based assessment*. New York, NY: Institute of Play. Retrieved from http://www.instituteofplay.org/work/projects/glasslab-research/

Mislevy, R. J., & Duran, R. P. (2014). A sociocognitive perspective on assessing EL students in the age of common core and Next Generation Science Standards. *TESOL Quarterly*, *48*, 560–585.

Mislevy, R. J., Haertel, G., Cheng, B.H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., M. Colker, A. M., Rutstein, D., & Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation, 19*, 121–140.

Mislevy, R. J., Riconscente, M. M., & Rutstein, D. W. (2009). *Design patterns for assessing model-based reasoning* (PADI-Large Systems Technical Report 6). Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

Moss, P. A., Pullin, D., Haertel, E. H., Gee, J. P., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. New York, NY: Cambridge University Press.

National Research Council (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas* (Committee on a Conceptual Framework for New K–12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education). Washington, DC: The National Academies Press.

NGSS Lead States. (2013a). *How to read the Next Generation Science Standards*. Retrieved January 23, 2016, from http://www.nextgenscience.org/sites/ngss/files/How%20to%20Read%20NGSS%20-%20Final%204-19-13.pdf

NGSS Lead States. (2013b). 4-ESS3-1 Earth and human activity. Retrieved January 23, 2016 from http://www.nextgenscience.org/4-ess3-1-earth-and-human-activity

Organisation for Economic Co-operation and Development. (2013, March). *PISA 2015 draft collaborative problem solving framework*. Retrieved January 19, 2016, from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Pandey, T. N. & Carlson, D. (1976). Assessing payoffs in the estimation of the mean using multiple matrix sampling designs. In D. N. M. de Gruijter & L. J. van der Kamp (Eds.) *Advances in psychological and educational measurement* (pp. 265–275). London, England: Wiley.

Perkins, D. N., & Salomon, G. (1989) Are cognitive skills context-bound? *Educational Researcher, 18*, 16–25

Resnick, L. B. (1994). Performance puzzles. *American Journal of Education, 102*, 511–526.

Robinson, P. (2010). Situating and distributing cognition across task demands: The SSARC model of pedagogic task sequencing. In M. Putz & L. Sicola (Eds.), *Cognitive processing in second language acquisition: Inside the learner's mind* (pp. 239–264). Amsterdam/Philadelphia PA: John Benjamins.

Roth, W.-M. (2009). Phenomenological and dialectical perspectives on the relation between the general and the particular. In K. Ercikan & W.-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 235–260). New York, NY: Routledge.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching, 33*, 1045–1063.

Ryans, D. G., & Frederiksen, N. (1951). Performance tests of educational achievement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 455–494). Washington, DC: American Council of Education.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved July 16, 2013, from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215–232.

Shepard, L. A. (2008). Formative classroom assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). New York, NY: Erlbaum.

Sikorski, T., & Hammer, D. (2010). A critique of how learning progressions research conceptualizes sophistication and progress. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 2010 International Conference of the Learning Sciences* (pp. 277–284). Chicago, IL: ISLS.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching*, *46*(6), 610–631.

Sottilare, R. A., Graesser, A. C., Hu, X, & Brawner, K. (Eds.) (2015). *Design recommendations for intelligent tutoring systems*. Orlando. FL: U.S. Army Research Laboratory.

Swain, M. (1985). Large-scale communicative language testing: A case study. In Y. P. Lee, A. C. Y. Y. Fok, R. Lord & G. Low (Eds.), *New directions in language testing* (pp. 35–46). Oxford, England: Pergamon.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296.

Toulmin, S. (1958). *The use of argument*. Cambridge: University Press.

Wertsch, J. (1994). The primacy of mediated action in sociocultural studies. *Mind, Culture, and Activity, 1*, 202–208.

Young, R. F. (2009). *Discursive practice in language learning and teaching*. Hoboken, NJ: Wiley-Blackwell.

Young, R. F., & He, A. W. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam and Philadelphia, PA: John Benjamins.